THE LOCAL BEHAVIOR OF THE SHRINK-WRAPPING ALGORITHM FOR LINEAR PROGRAMMING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Yuriy A. Zinchenko

January 2006

THE LOCAL BEHAVIOR OF THE SHRINK-WRAPPING ALGORITHM FOR LINEAR PROGRAMMING

Yuriy A. Zinchenko, Ph.D.

Cornell University 2006

Hyperbolic polynomials and their associated hyperbolicity cones have origins in partial differential equations. Recently, these structures have drawn considerable attention in the optimization community as well. It turns out that most of interior point methods (IPM) theory applies naturally to the class of conic programming problems arising from hyperbolicity cones. In particular, linear programming (LP), second-order conic programming (SOCP) and positive semi-definite programming (SDP) are themselves instances of conic programming problems of this kind.

The thesis consists of two parts. The first part is devoted to the structure of a particular family of hyperbolicity cones which give a sequence of relaxations to the nonnegative orthant. The second part contains analysis of the newly proposed algorithm for LP based on these relaxations.

While one can easily construct a logarithmic self-concordant barrier (SCB) functional for the hyperbolicity cone K_p associated to an arbitrary hyperbolic polynomial p, little is known about its dual cone K_{p_*}. This problem is closely related to LP itself: for the case of $p(x) = E_n(x) = \prod_{i=1\ldots n} x_i$ the (closure of the) hyperbolicity cone is self-dual and is, in fact, $\mathbb{R}^n_+$. Elementary symmetric polynomials can be thought of as derivative polynomials (in a certain sense) of $E_n(x)$, and are building blocks for hyperbolic polynomials themselves, their associated hyperbolicity cones

giving a natural sequence of relaxations for $\mathbb{R}^n_{++}$. We give an algebraic characterization for the dual cone associated with $p'(x) = E_{n-1}(x) = \sum_{1 \le i \le n} \prod_{j \neq i} x_j$ which was previously unknown and show how one can easily construct a SCB functional for this cone. We comment on possible extensions of this result.

Recently a new paradigm for LP has been proposed (J. Renegar). It relies on the consecutive relaxations of the nonnegative orthant using hyperbolicity cones associated with elementary symmetric functions. In a way this gives a generalization to the notion of a central path in IPM. We analyze the local behavior of the newly proposed algorithm in the neighborhood of the optimal LP solution demonstrating that the resulting sequence of iterates will converge at least super-linearly to the solution (under some non-degeneracy assumptions).

BIOGRAPHICAL SKETCH

Yuriy Zinchenko was born on April 16, 1975 in Kharkov, Ukraine. In May 1992 he graduated with honor (silver medal) from the high-school number 5 with specialization in physics and mathematics in Kharkov, Ukraine. In May 1996 Yuriy received his specialist degree with honor (red diploma) from the Kharkov State Polytechnic University, Kharkov, Ukraine.

In August 1999 Yuriy started his Ph.D. program at School of Operations Research and Industrial Engineering at Cornell University, Ithaca NY, USA, under the thesis guidance of Prof. James Renegar.

To my family.

ACKNOWLEDGEMENTS

My stay at Cornell has been a great experience. Firstly, I would like to thank two most important people that I met here, although they fall into quite distinct categories.

I would like to thank my fiance, Deniz Sezer, whom I met here at Cornell, for her love and support. I would like to express my deep gratitude to my advisor, Prof. James Renegar, for his guidance, infinite patience and support, and for always being straightforward.

I would like to thank Prof. Louis Billera and Prof. Sidney Resnick for serving on my committee.

The faculty at the department of ORIE makes this a terrific place to be. Especially, I would like to thank Prof. Gennady Samorodnitsky, Prof. David Shmoys, Prof. Robert Bland, Prof. Leslie Trotter, Prof. Shane Henderson and Prof. Michael Todd, for their help, compassion and encouragement.

I would like to thank Prof. Adrian Lewis for his comments on my thesis presentation and a couple of tennis matches that we had a chance to play. I would like to thank Prof. Leonard Gross (at the Department of Mathematics) for making taking courses at Cornell a very joyful experience and giving a good advice when it was needed.

I would like to thank all of my friends, Thanos Avramidis, Davina, Yulya, Mayur, Millie Chu, Dhruv, Tuncay, Amar, Chek Beng, Michael Wagner, Vardges, Alper, Jeorg, Stefan, Pascal, Peter, Oguzan, Retsef, Bharath, Nikolai, Sam Ehrlichman, Trevor Park, "Magic" Chandra, Van Anh, Sumit and others for turning my stay at Cornell into something I will always miss.

TABLE OF CONTENTS

LIST OF FIGURES

NOTATION

Let X denote a finite-dimensional real vector space.

For $x, y \in X, \langle x, y\rangle : X \times X \to \mathbb{R}$ – (a natural) inner product defined on a vector space X (e.g., for $x, y \in \mathbb{R}^n$, $\langle x, y\rangle = x^T y$).

For $Y \subseteq X$ we write cl$Y$ for the closure of Y (in norm-induced topology), int$Y$ for its interior, and ∂Y for its boundary.

We write $x \in \mathbb{R}^n$ for a column vector in $\mathbb{R}^n$ (component-wise referred to as $(x_1, x_2, \ldots, x_n)$ or $[x_1; x_2; \ldots; x_n]$, with its transpose $[x_1, x_2, \ldots, x_n]$ – a row vector).

For $x, y \in \mathbb{R}^n$,

- we say $x \leq y$ or $x < y$ if this inequality holds component-wise,
- we write x/y or $\frac{x}{y}$ for a vector in $\mathbb{R}^n$ whose components are x_i/y_i for all $i = 1, \ldots, n$,
- we write x^k for a vector whose components are the k^{th} power of the components of x.

More generally, for $x \in \mathbb{R}^n$ and a function $f : \mathbb{R} \to \mathbb{R}$ we write $f(x)$ or $(f(x_i))_{i=1}^n$ for a vector whose components are $f(x_i), i = 1, \ldots, n$.

For a vector $x \in \mathbb{R}^n$ and a linear subspace $L \subseteq \mathbb{R}^n$ we write x_L for the orthogonal projection of x onto L (i.e. $x_L = \text{proj}_L(x)$). Moreover, we write $x =_L y$ for $y \in \mathbb{R}^n$ if $x_L = y_L$.

For a vector $x \in \mathbb{R}^n$ and an arbitrary index $1 \leq i \leq n$, we write $x_{-i} \in \mathbb{R}^{n-1}$ for a vector whose i^{th} coordinate has been removed.

We write $\mathbf{1} \in \mathbb{R}^n$ ($\mathbf{0} \in \mathbb{R}^n$) for the vector whose components are all ones (zeros).

For a vector $x \in \mathbb{R}^n$ we write $[x]$ for Diag(x), the diagonal matrix with x along the diagonal.

Let $\mathbb{R}^n_+$ be the nonnegative orthant in $\mathbb{R}^n$ ($\{x \in \mathbb{R}^n : x \geq 0\}$).

Let $\mathbb{R}^n_{++}$ be the strictly positive orthant in $\mathbb{R}^n$ ($\{x \in \mathbb{R} : x > 0\}$).

For a matrix $A \in \mathbb{R}^{m \times n}$ we write null(A) for its null-space ($\{x \in \mathbb{R}^n : Ax = 0\}$) and range$(A)$ for its range ($\{x \in \mathbb{R}^n : \exists y \in \mathbb{R}^m, y^T A = x\}$)

For a matrix $A \in \mathbb{R}^{m \times n}$ we write $A_i (\in \mathbb{R}^m)$ for its i^{th} $(1 \leq i \leq n)$ column vector and $(A_i)_{i \in I}$, or just A_I, for a matrix in $\mathbb{R}^{m \times |I|}$ whose columns are the A_i columns of the original matrix A appearing in the exact order prescribed by I ($|I|$ is the cardinality of the index set I). In general we allow the elements in I to repeat (e.g. $I = \{1, 1, 2\}$). Likewise for a vector $x \in \mathbb{R}^n$ we write x_I for a vector whose components are $x_i, i \in I$.

Let $\mathbb{S}^k$ be the space of real symmetric $k \times k$ matrices.

We write $A \in \mathbb{S}^k, A \succ 0 (\succeq 0)$ if a matrix A is positive (semi-)definite.

Let $\mathbb{S}^k_+$ be the cone of positive semi-definite matrices ($\{A \in \mathbb{S}^k : A \succeq 0\}$).

Let $\mathbb{S}^k_{++}$ be the cone of positive definite matrices ($\{A \in \mathbb{S}^k : A \succ 0\}$).

Chapter 1

Introduction

1.1 Problem motivation

Letting X ($\equiv \mathbb{R}^n$) be equipped with an inner product $\langle \cdot, \cdot \rangle$, a *conic program* is an optimization problem of the form

$$(CP) \quad \{\inf_x \langle c, x \rangle : Ax = b, x \in K\}$$

with $K \subset \mathbb{R}^n$ being a closed convex cone (recall that a set is a cone if it is closed under multiplication by nonnegative reals), $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ and $A : \mathbb{R}^n \to \mathbb{R}^m$ being a linear operator ($A \in \mathbb{R}^{m \times n}$). It is well known that any convex optimization problem can be recast as conic programming problem.

The three most prominent instances of (CP) are:

- linear programming (LP): $X = \mathbb{R}^n, \langle x, y \rangle = x^T y, K = \mathbb{R}^n_+$ (the nonnegative orthant),
- second-order conic programming (SOCP): $X = \mathbb{R}^n, \langle x, y \rangle = x^T y, K = K_1 \times K_2 \times \cdots \times K_l$ with $K_i = \{(x, t) \in \mathbb{R}^{n_i - 1} \times \mathbb{R} : \|x\| \leq t\}$ ("second-order cones"; a.k.a., "Lorentz cones"), $\sum_{i=1}^{l} n_i = n$, and
- positive semi-definite programming (SDP): $X = \mathbb{S}^k$ (the space of real $k \times k$ symmetric matrices), $\langle x, y \rangle = x \circ y = \text{trace}(xy)$ and $K = \mathbb{S}^k_+$ (the cone of positive semi-definite matrices).

In applications, these three types of problems provide an extremely powerful modelling framework, ranging from production planning and relaxations for some combinatorial problems to control theory and polynomial programming [1], [2], [3],

[4], [5]. Also, they naturally arise as robust counterparts to one another in the presence of uncertainty in the initial data [2], [17].

Besides the modeling capabilities of these three types of problems, a key reason for their tremendous success is the existence of efficient algorithms to solve them; in particular, interior-point methods (IPM). In theory, any CP can be solved using IPM [6]. Moreover, LP, SOCP, SDP can be solved in polynomial time [6]. Despite promising theoretical results, however, the cost of the linear algebra involved remains a prohibitive factor in many applications (realistically we can hope to solve LP with $(m, n) \sim 10^6$ and SDP with $(m, n) \sim 10^3$). For extremely large scale problems, there is a marked need for algorithms more efficient than IPM.

The three types of problems are special cases of *hyperbolic programming*, a kind of convex optimization problem having rich algebraic structure. Hyperbolic polynomials and their associated hyperbolicity cones were first extensively studied in the context of partial differential equations. Recently, these structures have drawn considerable attention in the optimization community as well [10],[11]. It happens, for example, that most of richest IPM theory can be extended to hyperbolic programs.

Herein I address two problems:

(i) the structure of hyperbolicity cones associated with elementary symmetric polynomials and the structure of the dual cones,

(ii) the study of local convergence of the newly proposed "Shrink-Wrapping" algorithm for LP.

As an exemplary application of these optimization techniques I briefly describe the application of optimization to intensity modulated radiation therapy (IMRT) planning, a project I pursued in addition to my thesis research, in collaboration

with Prof. Shane Henderson (Cornell University), M. Chu (Cornell University) and Michael B. Sharpe (Princess Margaret Hospital, Toronto, Canada).

IMRT is one of the state of the art treatment techniques for cancer patients. During the course of the treatment high energy photons are delivered to a targeted area with very high precision. The objective is to deliver the prescribed uniform dose to the tumor while sparing the healthy surrounding tissues.

A commonly proposed approach is to use mixed integer programming to model the treatment process. The exact model poses a great computational challenge due to the large number of binary decision variables. To overcome this, one has to resort to the continuous relaxations of the model (for example, LP).

There are a few sources of uncertainty in the planning process which are typically overlooked: the displacement of the patient's body during a treatment session (usually the whole treatment is administered across multiple sessions), movement of the organs during the treatment, etc. If these small discrepancies in the initial data are ignored, it might result in accepting a seemingly good plan which turns out to be extremely unstable under the actual perturbations.

Under some very mild probabilistic assumptions we incorporate this uncertainty into our model using what is called a robust LP (which can be cast into SOCP, see[2],[17]). The proposed approach is feasible computationally and numerical results indicate its superiority over previous models.

I intend to continue working on these projects during my post-doctoral stay at McMaster University, Canada.

1.2 A brief review of the existing work

The extensive study of hyperbolic polynomials begins with the work of Lars Gårding (see [12]), which dates back to 1950's, in the context of partial-differential equations. In this work the author established a number of important results about the hyperbolic polynomials including the convexity of the associated hyperbolicity cones. The notion of hyperbolic programming was first introduced in [11]. In this work the author demonstrated, in particular, that the hyperbolic programming problems can be efficiently solved using the interior point methods, and gave a first characterization of the hyperbolicity cones as a set of polynomial inequalities (although quite different and more complicated then the one in [8] that we rely on). Further study of hyperbolic polynomials in the context of convex optimization was done by the group of authors of [10]. Here a number of important observations were made regarding the connections of hyperbolic polynomials with the symmetric functions, and in particular, the elementary symmetric functions. This work is an excellent introduction to hyperbolic polynomials in the context of mathematical programming. This line of research was continued in [8], where many important properties of the boundary of the hyperbolicity cones are revealed together with the relevance of the so-called hyperbolic derivative cones.

The SDP introduced above can be written as a constrained optimization problem with linear objective function subject to a finite number of polynomial inequalities. The hyperbolic programming problems mentioned above also admit similar representation. A question arises: "Is there additional similarity between them?".

It has been long hypothesized that the cone of positive semi-definite matrices and the hyperbolicity cones have a strong relationship. The most fundamental result in this area, bearing the name of Lax conjecture, (proposed by Peter Lax in

1958), was established only half a century later in [9]. The result was proven using the recent work in real-algebraic geometry presented in [15] on the representation of sets as linear matrix inequalities. In [14] a similar connection has been established for a quite broad family of hyperbolicity cones (the so-called homogeneous cones) with the cone of positive semi-definite matrices. Although the question about the representation of hyperbolicity cones as a system of polynomial inequalities has been resolved, little is known about the structure of the associated dual cones. The existence of a similar representation follows from work in real algebraic geometry on quantifier elimination, see, for example, [13]. Unfortunately it provides little or no insight into the problem, because of the immense computational complexity resulting from this approach.

An important connection regarding the SDP and general polynomially constrained optimization problems is established in [3] (see other related work by Parillo, [23]). The author draws the connection between these optimization problems and the well-studied problem of moments (on representation of the moments of some σ-finite measure in $\mathbb{R}^n$, [25]). The positivity conditions for polynomials in one variable are also studied in the fairly recent work of [26] and others in relationship to SDP. The earlier reference to polynomially constrained optimization problems can be found, for example, in the work of Shor, [24].

There is a vast amount of literature on the subject of the interior-point methods. I would like to point out the fundamental work of [6] together with an excellent and much more accessible exposition of this material in [7].

Analogously, there are many available sources on the applications of these optimization techniques (see, for example, [5], [4], [16], [1]). The work of [2] is a great guide to modeling capabilities of the conic programming problems.

Chapter 2

On the duals of a certain family of hyperbolicity cones

Definition 2.0.1. For a cone $K \subseteq \mathbb{R}^n$, the *dual cone* is defined as $K^* = \{y \in \mathbb{R}^n : \forall x \in K, \langle x, y\rangle \geq 0\}$

Often, the dual cone provides much information about the original CP (indeed, the most successful IPM algorithms are the so-called primal-dual algorithms, which follow the so-called central paths in K and K^* simultaneously). Hence, the understanding of the structure of both the primal cone and the dual cone for a given conic programming problem CP usually plays a very important role in achieving greater computational efficiency in solving these optimization problems.

We are concerned with what is called "hyperbolic programming". One of the questions we would like to answer is: "What can be said about the structure of the particular cones giving rise to these optimization problems?".

While a simple characterization for the hyperbolicity cones as a set of polynomial inequalities is known, little is known regarding the algebraic structure of their dual cones. That the dual cones can be represented by systems of polynomial inequalities follows from Tarski's establishment of quantifier elimination methods (see [13]). These methods, however, give little insight into the precise algebraic structure of the dual cones, because the methods result in extremely complicated systems of polynomial inequalities, even for hyperbolic polynomials in 3 variables.

It turns out that the question above is hard to answer in its full generality (at least no detailed answer has been given to it yet despite efforts by numerous re-

searchers, while some conjectures do exist). We will concentrate on a more specific question instead, which is also more relevant to the case of linear programming. We attempt to understand the structure of hyperbolicity cones associated with elementary symmetric polynomials (which is an important family of hyperbolicity cones) and the structure of the associated dual cones.

2.1 Hyperbolic programming

In what follows we introduce the notions of hyperbolic polynomials, the associated hyperbolicity cones and hyperbolic programming problem, together with some important properties that we will rely on later.

Definition 2.1.1. A nonconstant polynomial $p : X \to \mathbb{R}$ is *homogeneous of degree* m (m is a positive integer) if $p(tx) = t^m p(x)$, for all $t \in \mathbb{R}$ and every $x \in X$.

Definition 2.1.2. Suppose that $p : X \to \mathbb{R}$ is a homogeneous polynomial of degree m and $d \in X$ is such that $p(d) \neq 0$. Then p is *hyperbolic with respect to* d if the univariate polynomial $\lambda \mapsto p(x - \lambda d)$ has all roots real for every $x \in X$.

Examples:

- $X = \mathbb{R}^n$, $d = \mathbf{1} \in \mathbb{R}^n$. The n^{th} elementary symmetric function $p(x) = E_n(x) = \prod_{i=1}^n x_i$ is a hyperbolic polynomial with respect to d (for $E_n(x - \lambda\mathbf{1}) = \prod_{i=1}^n (x_i - \lambda)$ has roots x_i),
- $X = \mathbb{S}^k$ the space of real symmetric $k \times k$ matrices, $d = I \in \mathbb{S}^k$. The determinant $p(x) = \det(x)$ is a hyperbolic polynomial in direction d (for the eigenvalues of $x \in \mathbb{S}^k$ are the roots of $\det(x - \lambda I)$ and are real).

The roots are called the *eigenvalues* of x (in direction d), terminology motivated by the last example. We denote the eigenvalues by

$$\lambda_1(x) \leq \lambda_2(x) \leq \cdots \lambda_m(x)$$

Fact 2.1.3 (Gårding's [12])**.** *$\lambda_1(x)$ is a concave function of x.*

In fact if we introduce sums of the smallest k eigenvalues as follows

$$s_k := \sum_{i=1}^{k} \lambda_i$$

a more general (important) statement can be made

Fact 2.1.4 ([10])**.** *$s_k(x)$ is a concave function for any $k = 1, \ldots, m$.*

Definition 2.1.5. The *hyperbolicity cone of p* with respect to d, written $C(d)$ or $C(p, d)$, is the set $\{x \in X : p(x - \lambda d) \neq 0, \forall \lambda \leq 0\}$.

Note that $C(d) = \{x \in X : \lambda_1(x) > 0\}$.

Examples:

- $X = \mathbb{R}^n$, $d = \mathbf{1}$, $p(x) = E_n(x)$, then $C(d) = \mathbb{R}^n_{++}$,
- $X = \mathbb{S}^k$, $d = I$, $p(x) = \det(x)$, then $C(d) = \mathbb{S}^k_{++}$.

Lars Gårding ([12]) was the first one to study hyperbolicity cones carefully, in the early 1950's.

Given a hyperbolic polynomial p with respect to d the following are true:

Fact 2.1.6. *Given a pair p, d*

(i) $d \in C(d)$

(ii) *$C(d)$ is an open convex cone*

(iii) $clC(d) = \{x \in X : \lambda_1(x) \geq 0\}$

(iv) if $c \in C(d)$*, then* p *is hyperbolic in direction* c *and* $C(c) = C(d)$

Proof. We demonstrate that (ii) easily follows from 2.1.3. Indeed, suppose $x_1, x_2 \in X$ are both in $C(d)$, so that $\lambda_1(x_1) > 0$ and $\lambda_1(x_2) > 0$. Then for any $0 \leq \gamma \leq 1$, $\lambda_1((1-\gamma)x_1 + \gamma x_2) \geq (1-\gamma)\lambda_1(x_1) + \gamma\lambda_1(x_2) > 0$ by concavity of λ_1. For the rest of the proof, see[12]. □

Definition 2.1.7. A *hyperbolic programming program* is a CP where K is a closure of hyperbolicity cone.

Note that LP, SOCP, SDP are instances of hyperbolic programs.

2.2 Derivative polynomials and primal cone characterization

2.2.1 Derivative polynomials

Once we introduce hyperbolic polynomials a natural question might arise: can we construct new hyperbolic polynomials from the existing ones?

To give a partial answer, one can introduce a notion of the derivative of a hyperbolic polynomial p (with respect to d).

Suppose, as before, we have a hyperbolic polynomial p (of degree m) in direction d. Denote

$$p'(d, x) = \frac{\partial}{\partial t} p(x + td)|_{t=0} = \nabla_x p(x)^T d$$

We will refer to p' as the "derivative polynomial of p (with respect to d)" and usually will write $p'(x)$ instead of $p'(d, x)$ omitting (the parameter) d for simplicity

of notation (when the choice of d is obvious). By the root interlacing property for the polynomials with all real roots (by continuity between any two roots of $t \mapsto p(x+td)$ there is a root of $\frac{\partial}{\partial t}p(x+td)$) it follows that $p'(x)$ is also hyperbolic in direction d.

Similarly, (for a fixed hyperbolicity direction d) we can define higher derivatives $p'', p''', \ldots, p^{(m)}$. Note that since p was assumed to be of degree m, $p^{(m-1)}$ is linear and $p^m(x)$ is constant.

Examples:

- $X = \mathbb{R}^n$, $d = \mathbf{1}$, $p(x) = E_n(x)$, then

$$E_n^{(k)}(x) = (k!)E_{n-k}(x)$$

  where $E_j(x)$ is the j^{th} elementary symmetric function

$$E_1(x) = \textstyle\sum_{1\le i\le n} x_i, \quad E_2(x) = \sum_{1\le i<j\le n} x_i x_j, \quad \ldots, \quad E_n(x) = \prod_{1\le i\le n} x_i$$

- $X = \mathbb{R}^n$, $d \in \mathbb{R}^n_{++}$, $p(x) = E_n(x)$. Then by easy computation one can show that

$$E_n^{(k)}(x) = (k!)E_n(d)E_{n-k}\left(\left[\frac{x_1}{d_1}, \frac{x_2}{d_2}, \ldots, \frac{x_n}{d_n}\right]\right)$$

Remark 2.2.1. Note that in the second example we can choose any direction $d \in \mathbb{R}^n_{++}$ since $C(d) = C(c)$ for any $c \in C(d)$ (see 2.1.6), so the differentiation is well defined for any such c. Although $C(d) = C(c)$ for any $c \in C(d)$, the hyperbolicity cones corresponding to derivative polynomials $p', p'', \ldots$ might not necessarily coincide with one another. In the second example above in all the likelihood $C(p'(d,\cdot), d) \neq C(p'(c,\cdot), c)$ for $c \neq d$. This is an important observation to be made.

Remark 2.2.2. It should be noted that the elementary symmetric polynomials in the example above also play an important role in representing the derivative polynomials via the eigenvalues at a point $x \in X$. Namely, suppose we are given a hyperbolic polynomial p of degree m with respect to d. Corresponding to a point $x \in X$ we have m (not necessarily distinct) roots $\lambda_i(x), 1 \leq i \leq m$, of the univariate polynomial $\lambda \mapsto p(x - \lambda d)$, which are also exactly the minus roots of the polynomial $t \mapsto p(x + td)$. Now write $p(x + td) = \alpha \prod_{1 \leq i \leq m}(t + \lambda_i)$, and by homogeneity of p

$$\begin{aligned} p(d) &= \lim_{t\uparrow\infty} p\left(\frac{x}{t} + d\right) = \lim_{t\uparrow\infty} \frac{p(x+td)}{t^m} \\ &= \lim_{t\uparrow\infty} \alpha \frac{1}{t^m} \prod_{1\leq i\leq m} (t + \lambda_i) = \lim_{t\uparrow\infty} \alpha \prod_{1\leq i\leq m} \left(1 + \frac{\lambda_i}{t}\right) = \alpha \end{aligned}$$

so that we can write

$$p(x + td) = p(d) \prod_{1\leq i\leq m} (t + \lambda_i)$$

and consequently

$$\begin{aligned} p'(x) &= \frac{\partial}{\partial t} p(x + td)|_{t=0} = \frac{\partial}{\partial t} \left(p(d) \prod_{1\leq i\leq m} (t + \lambda_i) \right)_{t=0} \\ &= p(d) \sum_{1\leq i\leq m} \prod_{j\neq i} \lambda_i = p(d) E_{m-1}(\lambda) \end{aligned}$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$, and even more generally

$$p^{(k)}(x) = (k!) p(d) E_{m-k}(\lambda)$$

2.2.2 Cone characterization

It turns out that for pairs p, d, these derivative polynomials come in handy in characterizations of the (primal) hyperbolicity cone itself.

Denote $K_{p,d} := \mathrm{cl} C(p, d)$, the closure of hyperbolicity cone. When the choice of d is obvious, we will omit it from the notation, thus writing just K_p.

We present a well-known result giving one particular characterization of the (closure of) hyperbolicity cone (see, for example, [8]).

Theorem 2.2.3. *Suppose p is a hyperbolic polynomial of degree m with respect to d, $p(d) > 0$ (w.l.o.g.) and $p', p'', \ldots$ are defined as above. Then*

$$\begin{gathered} K_{p,d} = \{x \in \mathbb{R}^n : \\ p(x) \geq 0 \\ p'(x) \geq 0 \\ p''(x) \geq 0 \\ \vdots \\ p^{(m-1)}(x) \geq 0\} \end{gathered}$$

Proof. If $m = 1$ the result is trivial, so assume $m > 1$.

Denote $S := \{x \in \mathbb{R}^n : p(x) \geq 0, p'(x) \geq 0, p''(x) \geq 0, \ldots, p^{(m-1)}(x) \geq 0\}$. We show $S \subseteq K_{p,d}$ and $K_{p,d} \subseteq S$. Recall that $p^{(k)}(x) = (k!)p(d)E_{m-k}(\lambda)$ for $1 \leq k \leq (m-1)$.

$K_{p,d} \subseteq S$ inclusion: Suppose $x \in K_{p,d}$. Then all $\lambda_i(x) \geq 0$ and therefore $p^{(k)}(x) = (k!)p(d)E_{m-k}(\lambda) \geq 0$ for all $1 \leq k \leq (m-1)$.

$S \subseteq K_{p,d}$ inclusion by contradiction: Suppose it is not true, that is, suppose $\exists x \in X$ s.t. $x \in S$ but $x \notin K_{p,d}$. Since $x \notin K_{p,d}$ for the corresponding roots $\lambda(x)$ we must have

$$\lambda_1(x) \leq \lambda_2(x) \leq \cdots \leq \lambda_k(x) < 0 \leq \lambda_{k+1}(x) \leq \cdots \leq \lambda_m(x)$$

for some $1 \leq k \leq m$. By the root interlacing property the roots of p', denoted $\lambda'(x)$, will be located between the roots of p

$$\lambda_1(x) \leq \lambda'_1(x) \leq \lambda_2(x) \leq \cdots \leq \lambda_{m-1}(x) \leq \lambda'_{m-1}(x) \leq \lambda_m(x)$$

and so on for the higher derivatives of p, $p^{(k)}$, and the corresponding roots, $\lambda_i^{(k)}(x)$.

W.l.o.g. we can assume $\lambda_{k+1} > 0$, for if $\lambda(x) = 0$ is a root of multiplicity $1 \leq l < m$ (for the univariate polynomial $t \mapsto p(x - td)$) then obviously $p(x) = 0$ and we must also have

$$p'(x) = 0, p''(x) = 0, \ldots, p^{(l-1)}(x) = 0, p^{(l)}(x) \neq 0$$

Thus instead of $p(x)$ we can consider $p^{(l)}(x)$ with all the corresponding roots being distinct from 0.

Now, since none of $\lambda_i(x)$ are assumed to be 0, combined with the requirements for $x \in S$ this gives us $p(x) > 0$. Therefore there must be an even number of strictly negative roots $\lambda_i(x)$, that is k must be even, since $p(x) = p(d) \prod_{i=1}^{m}(\lambda_i(x))$.

Note that when taking a derivative polynomial $p'(x)$, by the root interlacing property it follows that only one root $\lambda_i'(x)$ is allowed to "cross 0" (that is we must have $\lambda_1'(x) \leq \cdots \leq \lambda_{k-1}'(x) < 0$ and $0 < \lambda_{k+1}'(x) \leq \cdots \leq \lambda_{m-1}'(x)$; only λ_k is sign-undetermined), but since we also request $p'(x) \geq 0$ we necessarily must have $\lambda_k'(x) \leq 0$. There can be only two case here.

Case 1: $\lambda_k'(x) = 0$ (note that now 0 can not be a multiple root for if it was, then it must also be a multiple root for $p(x)$). We can differentiate $p'(x)$ again and (by the same root interlacing) we must have $\lambda_1''(x) \leq \cdots \leq \lambda_{k-1}'' < 0$ while $0 < \lambda_k''(x) \leq \cdots \lambda_{m-2}''(x)$ thus giving us $p''(x) < 0$, so $x \notin S$, contradiction.

Case 2: $\lambda_k'(x) < 0$. If there are no positive roots left for $p'(x)$, then by differentiating it one step further the resulting polynomial must have an odd number of strictly negative roots, thus $p''(x) < 0$ and we have a contradiction. Alternatively (if $k < (m-1)$), we can differentiate $p'(x)$ one more time and we will arrive at cases 1 or 2 again but now applied to the roots $\lambda_i''(x)$ (with $\lambda_{k-1}''(x)$ sign-undetermined), so we can repeat the differentiation until a contradiction is established (obviously

we will have to do this at most $(m-2)$ times). □

Corollary 2.2.4. *Given a pair p, d we have the following cone inclusions:*

$$K_{p,d} \subseteq K_{p',d} \subseteq \cdots \subseteq K_{p^{(m-1)},d}$$

Proof. Just remove the corresponding polynomial inequalities from the description of $K_{p,d}$. □

Corollary 2.2.5. *Given a pair p, d, $p(d) > 0$, the boundary of $K_{p,d}$ satisfies*

$$\partial K_{p,d} = \{x \in \mathbb{R}^n : p(x) = 0, p'(x) \geq 0, \ldots, p^{(m-1)}(x) \geq 0\}$$

Proof. Follows from the root interlacing property for the polynomial with all real roots: the smallest eigenvalue of x with respect to (p, d), $\lambda_1(x)$, is the leftmost-most root of the corresponding univariate polynomial $t \mapsto p(x - td)$, and hence is the first one to cross 0 thus turning $p(x)$ into 0 as well. □

Proposition 2.2.6. *Suppose (for some $1 \leq r \leq (m-2)$) $x \in K_{p^{(r)},d}$ and $p^{(r+1)}(x) = 0$ (that is $x \in K_{p^{(r+1)},d}$). Then $x \in K_{p,d}$.*

Proof. By the root interlacing property for polynomials with all real roots it follows that $t = 0$ is a multiple root of $t \mapsto p^{(r)}(x + td)$ of multiplicity $l \geq 2$ (since $x \in K_{p^{(r)},d}$ all the roots of $t \mapsto p^{(r)}(x + td)$ must be non-positive. Also the roots of $t \mapsto p^{(r+1)}(x + td)$ must be interlaced with the roots of $p^{(r)}(x + td)$). Therefore, by the same interlacing property, 0 is a root of multiplicity $(l+1)$ for $t \mapsto p^{(r-1)}(x+td)$, and so on, until we get to p itself. Since 0 was the right-most root for $t \mapsto p^{(r)}(x+td)$ ($x \in K_{p^{(r)},d}$), it will also be the right-most root $t \mapsto p^{(0)}(x + td)$ (by counting the number of roots). So $x \in K_{p,d}$ (in fact $x \in \partial K_{p,d}$). □

In particular, for $X = \mathbb{R}^n$, $d \in \mathbb{R}^n_{++}$, $p(x) = E_n(x)$, then

$$\mathbb{R}^n_+ = K_{E_n,d} \subseteq K_{E_n^{(1)},d} \subseteq \cdots \subseteq K_{E_n^{(n-1)},d}$$

Note that $K_{E_n^{(n-1)},d}$ is just a half-space passing through the origin with normal vector $\mathbf{1}/d$. This gives us a natural sequence of relaxations of the nonnegative orthant, a pivotal observation for building the Shrink-Wrapping algorithm framework for linear programming.

For simplicity of notation we write $K_{k,d}$ for $K_{E_n^{(n-k)},d}$ to signify the degree of the underlying polynomial (recall that $E_n^{(n-k)}(x)$ will correspond to the k^{th} elementary symmetric polynomial evaluated at the "scaled" vector x/d, up to a constant multiplier given d fixed).

Although we have a simple algebraic characterization of the hyperbolicity cones, their dual cones are poorly understood, with some exceptions (e.g. [14]).

2.3 Semi-definite representability and the dual cones

It has been long hypothesized that the hyperbolicity cones and the cone of positive semi-definite matrices have strong connections. In particular one of the open questions is whether the hyperbolicity cones are more general than the linear sections of $\mathbb{S}^n_+$ (and consequently, whether hyperbolic programming is any more general than SDP).

In 1958, P. Lax conjectured that each hyperbolic polynomial $p(x)$ in 3 variables can be written as a determinant of a linear combination of three symmetric matrices, $A, B, C \in \mathbb{S}^d$, $p(x) = \det(x_1 A + x_2 B + x_3 C)$ (for some d), consequently each hyperbolicity cone in 3 variables can be realized as the intersection of $\mathbb{S}^d_+$ with an affine subspace of $\mathbb{S}^d$. The conjecture was recently established affirmatively in [9]

– as a corollary to work of J. William Helton and V. Vinnikov [15]. It remains open whether similar representations hold for hyperbolicity cones in more than three variables, although such representations have been established for important broad families of hyperbolicity cones (in particular, the so-called homogeneous cones,[14]).

It turns out that an SDP representation also explains the structure of the corresponding dual cones (under some mild assumptions). We now make this statement more precise.

Definition 2.3.1 (as in [2])**.** The (convex) set $X \subseteq \mathbb{R}^n$ is said to be *SDR (positive semi-definite representable)* if

$$x \in X \Leftrightarrow \mathcal{A}\begin{bmatrix} x \\ \\ u \end{bmatrix} + B \succeq 0, \text{ for some } u \in \mathbb{R}^m$$

where $B \in \mathbb{S}^k$ and $\mathcal{A} : \mathbb{R}^{n+m} \to \mathbb{S}^k$ can be written as

$$\mathcal{A}\begin{bmatrix} x \\ \\ u \end{bmatrix} = \sum_{i=1}^{n} x_i A_i + \sum_{j=1}^{m} u_j B_j$$

with $A_i, B_j \in \mathbb{S}^k$.

Fact 2.3.2. *If X is SDR then so is an affine image of X.*

Proof. Can easily show by switching to the appropriate basis in $\mathbb{S}^k$, see [2]. □

Definition 2.3.3. Let $X \subset \mathbb{R}^n$ be a convex set containing the origin. The *polar of X* is the set $X_* = \{y \in \mathbb{R}^n : y^T x \leq 1, \forall x \in X\}$.

In particular, the polar of a closed convex cone K is $-K^*$, minus the dual cone (easy to check). Polarity "nearly" preserves SDR.

Proposition 2.3.4 (An SDR analogue of the theorem about representability with second order cones, [2]). *Let $X \subset \mathbb{R}^n$ ($0 \in X$) be an SDR set:*

$$X = \{x : \exists u \ s.t. \ \mathcal{A}\begin{bmatrix} x \\ u \end{bmatrix} + B \succeq 0\}$$

Assume there exists $\bar{x}, \bar{u}$ such that $\mathcal{A}\begin{bmatrix} \bar{x} \\ \bar{u} \end{bmatrix} + B \succ 0$. Then the polar X is an SDR set

$$X_* = \{y : \exists \Lambda \succeq 0 \ s.t. \ \langle A_i, \Lambda\rangle = -y_i, i = 1 \dots n, \langle B_j, \Lambda\rangle = 0, j = 1 \dots m, \langle B, \Lambda\rangle \leq 1\}$$

Proof. Indeed, consider the SDP

$$(P(y)) \quad \inf_{x,u}\{-\begin{bmatrix} y \\ 0 \end{bmatrix}^T \begin{bmatrix} x \\ u \end{bmatrix} : \mathcal{A}\begin{bmatrix} x \\ u \end{bmatrix} + B \succeq 0 \ \}$$

then $y \in X_*$ iff $(P(y))$ is bounded below by -1. Since $(P(y))$ is strictly feasible, the Conic Duality Theorem implies these properties of $(P(y))$ will hold iff the dual problem

$$\sup_{\Lambda \succeq 0}\{\langle -B, \Lambda\rangle = -Tr(B\Lambda) : \langle A_i, \Lambda\rangle = -y_i, i = 1 \dots n, \langle B_j, \Lambda\rangle = 0, j = 1 \dots m\}$$

has a feasible solution with value of at least -1, that is,

$$X_* = \{y : \exists \Lambda \succeq 0 \text{ s.t. } \langle A_i, \Lambda\rangle = -y_i, i = 1 \dots n, \langle B_j, \Lambda\rangle = 0, j = 1 \dots m, \langle B, \Lambda\rangle \leq 1\}$$

□

Corollary 2.3.5 (Dual cone of an SDR cone). *If $K \subset \mathbb{R}^n$ is a (closed) cone with nonempty interior and*

$$K = \{x \in \mathbb{R}^n : \exists u \ s.t. \ \mathcal{A}\begin{bmatrix} x \\ u \end{bmatrix} + B \succeq 0\}$$

then its dual satisfies

$$K^* = \{y \in \mathbb{R}^n : \exists \Lambda \ s.t. \ \begin{pmatrix} y \\ 0 \end{pmatrix} = \mathcal{A}^*\Lambda, \langle B, \Lambda \rangle \leq 0, \Lambda \succeq 0\}$$

where $\mathcal{A}^* : \mathbb{S}^k \to \mathbb{R}^{n+m}$ *is the adjoint of* $\mathcal{A}$*, defined as*

$$\mathcal{A}^*\Lambda = (\langle A_1, \Lambda \rangle, \langle A_2, \Lambda \rangle, \cdots, \langle A_n, \Lambda \rangle, \langle B_1, \Lambda \rangle, \cdots, \langle B_m, \Lambda \rangle)^T$$

Proof. Considering the primal-dual pair

$$\inf_{x,u} \{ \begin{bmatrix} y \\ 0 \end{bmatrix}^T \begin{bmatrix} x \\ u \end{bmatrix} : \mathcal{A} \begin{bmatrix} x \\ u \end{bmatrix} + B \succeq 0 \ \}$$

and

$$\sup_{\Lambda \succeq 0} \{\langle -B, \Lambda \rangle = -Tr(B\Lambda) : \langle A_i, \Lambda \rangle = y_i, i = 1 \ldots n, \langle B_j, \Lambda \rangle = 0, j = 1 \ldots m\}$$

we conclude that $y \in K^*$, iff the first problem is bounded below by 0, and hence iff the second has a feasible solution with the value of at least 0. Thus

$$K^* = \{y : \exists \Lambda \succeq 0 \text{ s.t. } \langle A_i, \Lambda \rangle = y_i, i = 1 \ldots n, \langle B_j, \Lambda \rangle = 0, j = 1 \ldots m, \langle B, \Lambda \rangle \leq 0\}$$

which can be rewritten as above. □

2.4 Elementary symmetric polynomials and the ratio functional

For a triple p, d, p' we now introduce the ratio functional $q_d := p/p'$ which has a nice geometric property, namely, it is concave on $K_{p',d}$. This functional will not only help us to analyze the set of necessary conditions for a point $x \in \mathbb{R}^n$ to belong to a hyperbolicity cone corresponding to the k^{th} elementary symmetric polynomial E_k, but will also prove to be important in understanding the new algorithmic setting for linear programming (the Shrink-Wrapping algorithm).

Proposition 2.4.1 (Concavity of the quotient functional in the special case where $q_n(x) = E_n/E_{n-1}(x)$)**.** *Let $x \in \mathbb{R}^n, d \equiv \mathbf{1} \in \mathbb{R}^n$ and $p(x) = \prod_{i=1}^n x_i$. Let $p'(x) = \mathbf{1}^T \nabla_x p = \sum_{i=1}^n \prod_{j \neq i} x_i$ be the derivative polynomial of p (in the direction $d \equiv \mathbf{1}$). Let K_p and $K_{p'}$ be the hyperbolicity cones corresponding to p and p' respectively. Denote*

$$q_n(x) := \frac{p(x)}{p'(x)}$$

Then $q_n(x)$ is concave over $K_{p'}$.

Proof. We proceed by evaluating the Hessian of $q_n(x)$. Note that

$$q_n(x) = \frac{p(x)}{p'(x)} = \frac{1}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}$$

so

$$\nabla q_n(x) = \nabla \left(\frac{1}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}} \right)$$

and

$$\frac{\partial q_n}{\partial x_i}(x) = \frac{-1}{\left(\frac{1}{x_1} + \cdots + \frac{1}{x_n}\right)^2} \frac{-1}{(x_i)^2} = \frac{\left(\frac{1}{x_i}\right)^2}{\left(\frac{1}{x_1} + \cdots + \frac{1}{x_n}\right)^2}$$

for $i = 1..n$. In order to evaluate the Hessian $\nabla^2 q_n(x)$, differentiate the expression above again. For $i \neq j$:

$$\begin{aligned} \frac{\partial^2 q_n(x)}{\partial x_i \partial x_j} &= \left(\frac{1}{x_i}\right)^2 \frac{-2}{\left(\frac{1}{x_1} + \cdots + \frac{1}{x_n}\right)^3} (-1) \left(\frac{1}{x_j}\right)^2 \\ &= \frac{2\left(\frac{1}{x_i}\right)^2 \left(\frac{1}{x_j}\right)^2}{\left(\frac{1}{x_1} + \cdots + \frac{1}{x_n}\right)^3} = \frac{2\left(\frac{1}{x_1} + \cdots + \frac{1}{x_n}\right)\left(\frac{1}{x_i}\right)^2 \left(\frac{1}{x_j}\right)^2}{\left(\frac{1}{x_1} + \cdots + \frac{1}{x_n}\right)^4} \end{aligned}$$

so

$$(\nabla^2 q_n)_{ij} = \left(\nabla q_n \left(2I \left(\frac{1}{x_1} + \cdots + \frac{1}{x_n} \right) \right) \nabla {q_n}^T \right)_{ij}$$

where I is the identity matrix.

Similarly, for $i = j$:

$$\begin{aligned}\frac{\partial^2 q_n(x)}{\partial x_i \partial x_i} &= \frac{2\left(\frac{1}{x_1}+\cdots+\frac{1}{x_n}\right)\left(\frac{1}{x_i}\right)^2\left(\frac{1}{x_i}\right)^2}{\left(\frac{1}{x_1}+\cdots+\frac{1}{x_n}\right)^4} + (-2)\frac{\left(\frac{1}{x_i}\right)^3}{\left(\frac{1}{x_1}+\cdots+\frac{1}{x_n}\right)^2} \\ &= \frac{2\left(\frac{1}{x_i}\right)^2\left(\frac{1}{x_i}\right)^2}{\left(\frac{1}{x_1}+\cdots+\frac{1}{x_n}\right)^4}\left(\left(\frac{1}{x_1}+\cdots+\frac{1}{x_n}\right) - \frac{\left(\frac{1}{x_1}+\cdots+\frac{1}{x_n}\right)^2}{\left(\frac{1}{x_i}\right)}\right)\end{aligned}$$

We can rewrite the Hessian of q_n in the following form:

$$\begin{aligned}\nabla^2 q_n &= 2\left(\frac{1}{x_1}+\cdots+\frac{1}{x_n}\right)\left(\nabla q_n \nabla q_n{}^T - \mathrm{Diag}\left(\left(\frac{\left(\frac{1}{x_i}\right)^4\left(\frac{1}{x_1}+\cdots+\frac{1}{x_n}\right)x_i}{\left(\frac{1}{x_1}+\cdots+\frac{1}{x_n}\right)^4}\right)^n_{i=1}\right)\right) \\ &= 2\left(\frac{1}{x_1}+\cdots+\frac{1}{x_n}\right)\left(\nabla q_n \nabla q_n{}^T - \mathrm{Diag}\left(\left(\frac{\left(\frac{1}{x_i}\right)^3}{\left(\frac{1}{x_1}+\cdots+\frac{1}{x_n}\right)^3}\right)^n_{i=1}\right)\right)\end{aligned}$$

Note that $q_n(x) < 0$ in the interior of $K_{p'} \backslash K_p$ since one of the roots of p becomes negative (all of the roots of p' are positive though). Since $\frac{1}{x_1}+\cdots+\frac{1}{x_n} = \frac{1}{q_n(x)}$, $\frac{1}{x_1}+\cdots+\frac{1}{x_n} < 0$ and amongst terms of the form $\frac{1}{x_i}$ exactly one of them is negative while all the rest are positive.

Denote

$$M = -\mathrm{Diag}\left(\left(\frac{\left(\frac{1}{x_i}\right)^3}{\left(\frac{1}{x_1}+\cdots+\frac{1}{x_n}\right)^3}\right)^n_{i=1}\right)$$

Thus M has all diagonal entries > 0 with the exception of one (also note that all these diagonal entries are precisely the eigenvalues for M). By adding a rank-one positive semi-definite matrix $\nabla q_n \nabla q_n{}^T$ to M the eigenvalues of a resulting matrix can only be shifted to the right (i.e. become "more positive"). But note that $\nabla^2 q_n(x)x = 0$ (because q_n is homogeneous of degree 1), henceforth the smallest (i.e. the negative) eigenvalue of M had become 0 under this perturbation (the addition of rank-one matrix $\nabla q_n \nabla q_n{}^T$). Hence $\nabla^2 q_n(x) \preceq 0$ (as a negative multiple of a positive semi-definite matrix analyzed above).

As of $q_n(x)$ being concave inside K_p ($\mathbb{R}^n_{++}$), one can show this directly. Recall that

$$q_n(x) = \frac{p(x)}{p'(x)} = \frac{1}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}$$

We need to show that

$$q_n(\alpha x + (1-\alpha)y) \geq \alpha q_n(x) + (1-\alpha) q_n(y) \ , \ \forall \alpha \in [0,1], x, y \in \mathbb{R}^n_{++}$$

In other words, we need to show that

$$\begin{aligned} A_1 &:= \frac{1}{\frac{1}{\alpha x_1 + (1-\alpha)y_1} + \cdots + \frac{1}{\alpha x_n + (1-\alpha)y_n}} \\ &\geq \frac{1}{\frac{1}{\alpha x_1} + \cdots + \frac{1}{\alpha x_n}} + \frac{1}{\frac{1}{(1-\alpha)y_1} + \cdots + \frac{1}{(1-\alpha)y_n}} =: A_2 \end{aligned}$$

Note that $\frac{1}{x}$ is convex on $\mathbb{R}_{++}$, i.e., $\frac{\alpha}{x} + \frac{(1-\alpha)}{y} \geq \frac{1}{\alpha x + (1-\alpha)y}$, from which follows that

$$\begin{aligned} A_1 &\geq \frac{1}{\frac{\alpha}{x_1} + \frac{(1-\alpha)}{y_1} + \cdots + \frac{\alpha}{x_n} + \frac{(1-\alpha)}{y_n}} \\ &= \frac{1}{\alpha\left(\frac{1}{x_1} + \cdots + \frac{1}{x_n}\right) + (1-\alpha)\left(\frac{1}{y_1} + \cdots + \frac{1}{y_n}\right)} \\ &\geq \frac{\alpha}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}} + \frac{(1-\alpha)}{\frac{1}{y_1} + \cdots + \frac{1}{y_n}} = A_2 \end{aligned}$$

□

Remark 2.4.2. The result that $q_n(x) = \frac{p(x)}{p'(x)} = \frac{E_n(x)}{E_{n-1}(x)}$ is concave on $K_p = \mathbb{R}^n_{++}$ also follows from Theorem 3.8 in [10]. In fact that theorem establishes an even more general result for arbitrary p and p' (the result is that if $q(y)$ is a symmetric convex function on $\mathbb{R}^m_+$ and $\lambda(x)$ is a function mapping from $\mathbb{R}^n$ onto the roots of some hyperbolic polynomial of degree m then the composite map $q \circ \lambda$ is convex on K_p). The result established above extends the domain of convexity/concavity to $K_{p'}$ for a particular function $q_n(x)$. Also note that if L is an affine space not containing the origin then $q_n(x)$ is strictly concave on the relative interior of $(K_{p'} \setminus K_p) \bigcap L$)

since $\nabla^2 q_n(x) \prec 0$ on this set because we eliminate the only possible direction of singularity for this matrix.

Theorem 2.4.3 (Concavity of the ratio functional for the elementary symmetric polynomials)**.** *Assume $2 \leq k \leq n$. Let $d \in \mathbb{R}^n_{++}$, $p(x) = E^k_n(x)$ and $p' = E^{(k+1)}_n(x)$ (with respect to d). Then*

$$q_d(x) := \frac{p(x)}{p'(x)}$$

is concave over $K_{p'}$.

Proof. We will proceed by considering $q_d(x) = \frac{p(x)}{p'(x)}$ as a composite function $q_d = \Psi \circ \Phi$ where $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ maps $\mathbb{R}^n$ onto the roots $\lambda_j(x)$ (in increasing order) of the polynomial $t \mapsto p(x - td)$ and $\Psi : \mathbb{R}^m \to \mathbb{R}$ maps these roots onto the range of q_d. Recall that the value of p at a point x can be written as $p(x) = p(d) \prod_{j=1}^m \lambda_j(x)$ and $p'(x) = p(d) \sum_{j=1}^m \prod_{k \neq j} \lambda_k(x)$. If we let $\Psi(y) := \frac{E_m(y)}{E_{m-1}(y)}$ (defined for $y \in C(E_{m-1}, \mathbf{1}) \subset \mathbb{R}^m$), then we can write $q_d(x) = \frac{p(x)}{p'(x)} = \Psi(\Phi(x))$.

If $\Phi(x)$ is differentiable at a point $x \in \text{int}K_{p'}$, we can express the gradient and the Hessian of $q_d(x)$ as follows:

$$\begin{aligned} \nabla q_d(x) &= \left(\frac{\partial q(x)}{\partial x_i} \right)^n_{i=1} = \left(\frac{\partial \Psi(\Phi(x))}{\partial x_i} \right)^n_{i=1} \\ &= \left(\sum_{j=1}^m \frac{\partial \Psi}{\partial \Phi_j} \frac{\partial \Phi_j}{\partial x_i} \right)^n_{i=1} = \Phi'(x)^T \nabla \Psi(\Phi(x)) \end{aligned}$$

$$\nabla^2 q_d = (\Phi')^T \nabla^2 \Psi \Phi' + \sum_{k=1}^n \frac{\partial \Psi}{\partial \Phi_k} \left[\frac{\partial^2 \Phi_k}{\partial x_i \partial x_j} \right]^n_{i,j=1}$$

where

$$\Phi'(x) = \left[\frac{\partial \Phi_i}{\partial x_j} \right]^n_{i,j=1}$$

(the Jacobian of $\Phi(x)$). Even though $\Phi(x)$ is not differentiable everywhere in its domain, it is differentiable at all x such that the components of $\Phi(x)$, $\lambda_i(x)$, are

distinct (see, for example, [22]). For $p(x) = E_n^k(x)$ (with respect to hyperbolicity direction d), if x is not a such point, i.e., if $\exists m > 1$ such that $\lambda_j = \cdots = \lambda_{j+m}$ for some j, then by the root interlacing property it must be that $x_j/d_j = \cdots = x_{j+k+m}/d_{j+k+m}$, since these are the roots of $t \mapsto E_n(x - td)$. We can easily choose a different hyperbolicity direction $d' \in B(d, \epsilon) \subset \mathbb{R}^n_{++}$ (in a small ball of radius ϵ around d) such that the roots of $t \mapsto E_n(x - td')$ (i.e., x_i/d'_i) will be distinct, and so will be the roots of $t \mapsto E_n^{(k)}(x - td)$. Since q_d is known to be differentiable at any $x \in \text{int}K_{p'}$, we can use the limiting argument letting $\epsilon \downarrow 0$ (so that $\nabla q_{d'}(x) \to \nabla q_d(x)$ and $\nabla^2 q_{d'}(x) \to \nabla^2 q_d(x)$ as $d' \to d$). In case if x is a point where $\Phi(x)$ is non-differentiable, the equality above should be taken in the limiting sense.

For the Hessian, the first term in the expression above is negative semi-definite (since $\nabla^2 \Psi \preceq 0$ on $C(E_{n-1}, \mathbf{1})$ that corresponds to $\text{int}K_{p'}$, see Proposition 2.4.1 above). The potential problem comes with the second term. To fix this, we consider two slightly different functions Φ^s and Ψ^s.

Denote by A the linear map $\lambda \mapsto s$ where $s_1 = \lambda_1, s_2 = \lambda_1 + \lambda_2, \ldots, s_n = \lambda_1 + \cdots + \lambda_n$ (s_k is the sum of smallest k eigenvalues λ_j). Clearly, A is invertible, and if we introduce $\Psi^s := \Psi \circ A^{-1}, \Phi^s := A \circ \Phi$, then $q_d(x) = \Psi^s(\Phi^s(x))$.

Observe that $\nabla^2 \Psi^s \preceq 0$ on $A\,(C(E_{m-1}, \mathbf{1}))$. Also, $\nabla^2 \Phi_k^s(x) = \left[\frac{\partial^2 \Phi_k^s}{\partial x_i \partial x_j}\right]_{i,j=1}^n \preceq 0$ since the functions $\Phi_k^s(x)$ are concave (see Fact 2.1.4). We show that $\frac{\partial \Psi^s}{\partial s_k} \geq 0$ for all k (recall $\nabla^2 q_d = ((\Phi^s)')^T \nabla^2_{ss} \Psi^s (\Phi^s)' + \sum_{k=1}^n \frac{\partial \Psi^s}{\partial s_k} \left[\frac{\partial^2 \Phi_k^s}{\partial x_i \partial x_j}\right]_{i,j=1}^n$).

Recall that

$$\frac{\partial \Psi}{\partial \lambda_i} = \frac{-1}{(\lambda_i)^2} \frac{-1}{\left(\frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_m}\right)^2} \geq 0$$

and

$$\begin{array}{cc} s_1 = \lambda_1 & \lambda_1 = s_1 \\ \vdots & \vdots \\ s_i = \lambda_1 + \cdots + \lambda_i & \lambda_i = s_i - s_{i-1} \\ \vdots & \vdots \\ s_m = \lambda_1 + \cdots + \lambda_m & \lambda_m = s_m - s_{m-1} \end{array}$$

then for $i = m$

$$\frac{\partial \Psi^s}{\partial s_m} = \frac{\partial (s_m - s_{m-1})}{\partial s_m} \frac{\partial \Psi}{\partial \lambda_m} = \frac{\partial \Psi}{\partial \lambda_m} \geq 0$$

and for $i < m$

$$\begin{aligned} \frac{\partial \Psi^s}{\partial s_i} &= \frac{\partial (s_i - s_{i-1})}{\partial s_i} \frac{\partial \Psi}{\partial \lambda_i} + \frac{\partial (s_{i+1} - s_i)}{\partial s_i} \frac{\partial \Psi}{\partial \lambda_{i+1}} \\ &= \frac{1}{(\lambda_i)^2} \frac{1}{\left(\frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_m}\right)^2} - \frac{1}{(\lambda_{i+1})^2} \frac{1}{\left(\frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_m}\right)^2} \end{aligned}$$

Note that since on K_p we have $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m$, it follows that $\frac{\partial \Psi^s}{\partial s_i} \geq 0, \forall i$, on K_p. We need a larger domain, namely $K_{p'}$.

On $K_{p'}$ we have $0 \leq \lambda_2 \leq \cdots \leq \lambda_m$ and thus $\frac{\partial \Psi^s}{\partial s_i} \geq 0, \forall i \geq 2$. On $K_{p'} \backslash K_p$ we have $\lambda_1 \leq 0 \leq \lambda_2 \leq \cdots \leq \lambda_m$, but we also must have $|\lambda_1| \leq |\lambda_2|$ (see the following lemma and the corollary). Therefore $\frac{\partial \Psi^s}{\partial s_i} \geq 0, \forall i$, for any x in $K_{p'}$, which completes the proof. □

Lemma 2.4.4. *Let $p(t)$ be an arbitrary polynomial over $\mathbb{R}$ of degree m with all distinct real roots and let p' be its derivative polynomial. Let λ and ξ be the roots of p and p' with $\lambda_1 < \xi_1 < \lambda_2 < \xi_2 < \cdots < \xi_{m-1} < \lambda_m$ ($p(t) = \alpha \prod_{i=1}^{m} (t - \lambda_i), p'(t) = \alpha \sum_{i=1}^{m} \prod_{j \neq i} (t - \lambda_i) = \alpha \prod_{k=1}^{m-1} (t - \xi_k)$). Then ξ_i satisfies*

$$\frac{1}{\lambda_1 - \xi_i} + \frac{1}{\lambda_2 - \xi_i} + \cdots + \frac{1}{\lambda_m - \xi_i} = 0$$

and

$$\xi_i \in \left[\lambda_i + \frac{(\lambda_{i+1} - \lambda_i)}{m - i + 1}, \lambda_{i+1} - \frac{(\lambda_{i+1} - \lambda_i)}{i + 1} \right]$$

Proof. Without loss of generality assume p is monic. Consider the ratio $\frac{p'}{p}$ on the real line except for the points where $p(t) = 0$ (in order to be well defined). Observe that $\frac{p'(t)}{p(t)} = 0$ iff $p'(t) = 0$ on this set. But

$$\frac{p'(t)}{p(t)} = \frac{\sum_{i=1}^{m} \prod_{j \neq i} (t - \lambda_i)}{\prod_{i=1}^{m} (t - \lambda_i)} = -\left(\frac{1}{\lambda_1 - t} + \cdots + \frac{1}{\lambda_m - t} \right)$$

hence the first part of the statement follows.

In order to get the bounds on ξ_i consider the equation above being set to 0 and note that

$$\begin{aligned} -\frac{1}{\lambda_i - \xi_i} &\leq -\left(\frac{1}{\lambda_1 - \xi_i} + \cdots + \frac{1}{\lambda_i - \xi_i} \right) \\ &= \left(\frac{1}{\lambda_{i+1} - \xi_i} + \cdots + \frac{1}{\lambda_m - \xi_i} \right) \leq \frac{m - i}{\lambda_{i+1} - \xi_i} \end{aligned}$$

and

$$\begin{aligned} -\frac{i}{\lambda_i - \xi_i} &\geq -\left(\frac{1}{\lambda_1 - \xi_i} + \cdots + \frac{1}{\lambda_i - \xi_i} \right) \\ &= \left(\frac{1}{\lambda_{i+1} - \xi_i} + \cdots + \frac{1}{\lambda_m - \xi_i} \right) \geq \frac{1}{\lambda_{i+1} - \xi_i} \end{aligned}$$

which follows from the ordering of the roots. Then

$$\lambda_i + \frac{(\lambda_{i+1} - \lambda_i)}{m - i + 1} \leq \xi_i \leq \lambda_{i+1} - \frac{(\lambda_{i+1} - \lambda_i)}{i + 1}$$

□

Corollary 2.4.5. *Let p and p' be an arbitrary hyperbolic polynomial (of degree m) and its derivative (w.r.t. d), and let $K_p, K_{p'}$ be the corresponding cones. Let $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m) \in \mathbb{R}^m$ with $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m$ be the roots of $t \mapsto p(x - td)$. Then $|\lambda_1| \leq |\lambda_2|$ on $K_{p'} \backslash K_p$.*

Proof. Denote the roots of p' by $\xi_1 \leq \xi_2 \leq \ldots \leq \xi_{m-1}$. From the previous lemma we have that $\xi_1 \leq \lambda_2 - \frac{(\lambda_2 - \lambda_1)}{2}$. But we know that $\xi_1 \geq 0$ (since we are in $K_{p'}$).

Therefore $\frac{\lambda_2}{2} + \frac{\lambda_1}{2} \geq \xi_1 \geq 0$ and $\lambda_2 \geq -\lambda_1$ and since $\lambda_1 \leq 0$ on $K_{p'} \backslash K_p$ we are done. $\square$

The corollary completes the proof of Theorem 2.4.3.

Remark 2.4.6 (On generalization of Theorem 2.4.3). I hypothesize that a similar argument can be applied to the case of an arbitrary hyperbolic polynomial p and its derivative polynomial p' (with respect to some d). The only part that requires augmentation in the proof is the justification of differentiability assumption for Φ. This goes beyond the scope of our interest here and therefore has not been explored.

2.5 On the structure of the associated hyperbolicity cones and their dual cones

In $\mathbb{R}^n$ the elementary symmetric polynomials[1] can be thought of as derivative hyperbolic polynomials (with respect to $\mathbf{1}$) of the product of all the coordinates of $x \in \mathbb{R}^n$, $E_n(x)$, and thus are hyperbolic polynomials themselves. The associated hyperbolicity cones for $k = 1, \ldots n-1$ give a natural sequence of relaxations to the nonnegative orthant $\mathbb{R}^n_+ \equiv K_{E_n,\mathbf{1}} \subset K_{E_{n-1},\mathbf{1}} \subset \cdots \subset K_{E_1,\mathbf{1}}$.

Following an observation that $K_{E_k,\mathbf{1}}$ has a recursive structure similar to $\mathbb{R}^n_+ \equiv K_{E_n,\mathbf{1}}$, to gain insight into the dual cones $K^*_{E_k,\mathbf{1}}$, we create a suitable decomposition of the cone $K_{E_{n-1},\mathbf{1}}$ into smaller convex cones, suitable in the sense that each of the smaller cones admits a positive semi-definite representation (see the definition 2.3.1). Relying on the duality theory for SDP, we then obtain the dual cone for each of the smaller cones as an SDR set in itself, and finally, we reconstruct $K^*_{E_{n-1},\mathbf{1}}$

[1] $E_k = \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} \prod_{j=1}^{k} x_{i_j}$

as the intersection.

This result not only gives an easy characterization of the dual cone to $K_{E_{n-1},\mathbf{1}}$ (in particular, an SDR representation), but also reveals the geometry of both the primal and the dual cone in this particular case. Some concluding remarks on the possible extension of this result and its implications are made at the end of this section.

Throughout this section fix the underlying vector space to be $\mathbb{R}^n$, the hyperbolicity direction $d \equiv \mathbf{1}$. For a fixed $2 \leq k \leq (n-1)$, denote $p(t) : t \mapsto E_k(x + t\mathbf{1})$, for $1 \leq i \leq n$ denote $p_{-i}(t) : t \mapsto E_k(x_{-i} + t\mathbf{1}_{-i})$ and $p'_{-i}(t) : t \mapsto (n-k)E_{k-1}(x_{-i} + t\mathbf{1}_{-i})$.

2.5.1 The recursive structure of the hyperbolicity cones for elementary symmetric polynomials

Observe the recursive expression $E_k(x) = x_i E_{k-1}(x_{-i}) + E_k(x_{-i})$ for any $n > k \geq 2$ and an arbitrary index i, where $E_k(\cdot_{-i}) : \mathbb{R}^{n-1} \to \mathbb{R}$ is the k^{th} elementary symmetric function on $\mathbb{R}^{n-1}$.

Theorem 2.5.1 (Necessary condition for $x \in K_{E_k}$)**.** *Assume $2 \leq k \leq n$. Then $x \in K_{E_k(\cdot)}$ only if $x_{-i} \in K_{E_{k-1}(\cdot_{-i})}$, $\forall i$.*

Proof. If $k = n$ the result is obvious, so assume $k < n$. Fix i. Write $E_k(x) = x_i E_{k-1}(x_{-i}) + E_k(x_{-i})$ and recall that $x \in K_{E_k(\cdot)}$ iff $p(t) : t \mapsto E_k(x + t\mathbf{1})$ has only non-positive roots. $E_k(\cdot_{-i})$ and $E_{k-1}(\cdot_{-i})$ are both hyperbolic along $\mathbf{1}_{-i} \in \mathbb{R}^{n-1}$ and

$$\lim_{t\uparrow\infty} \tfrac{E_{k-1}(x_{-i}+t\mathbf{1}_{-i})}{t^{k-1}} \geq 0 \quad \text{and} \quad \lim_{t\uparrow\infty} \tfrac{E_k(x_{-i}+t\mathbf{1}_{-i})}{t^k} \geq 0$$

for $\forall x \in \mathbb{R}^n$ (as $t \uparrow \infty$ both $E_{k-1}(x_{-i} + t\mathbf{1}_{-i})$ and $E_k(x_{-i} + t\mathbf{1}_{-i})$ will eventually be ≥ 0).

Using $p_{-i}(t), p'_{-i}(t)$ as defined previously we can write $p(t) = E_k(x + t\mathbf{1}) = \frac{(x_i+t)}{n-k} p'_{-i}(t) + p_{-i}(t)$.

Suppose $x_{-i} \notin K_{E_{k-1}(\cdot_{-i})}$, so there must be at least one positive root of $p'_{-i}(t)$. We also know that roots of $p_{-i}(t)$ and $p'_{-i}(t)$ are interlaced: enumerating all roots (including multiplicities) of $p_{-i}(t)$ as $\{t_i : i = 1, \ldots, k\}$ and roots of $p'_{-i}(t)$ as $\{t'_i : i = 1, \ldots, (k-1)\}$ in non-decreasing order we must have $t_1 \leq t'_1 \leq t_2 \leq t'_2 \leq \cdots \leq t_{k-1} \leq t'_{k-1} \leq t_k$, $0 < t'_{k-1} \leq t_k$ and also from the observation made about signs of $p_{-i}(t)$ and $p'_{-i}(t)$ as $t \uparrow \infty$ we get that

$$p'_{-i}(t) \geq 0 \text{ for } t \geq t'_{k-1},$$
$$p_{-i}(t'_{k-1}) \leq 0 \text{ and } p_{-i}(t) \geq 0 \text{ for } t \geq t_k$$

We consider three cases depending on the value x_i.

Case 1. Suppose that $-x_i \leq t'_{k-1}$. Then

$$p(t'_{k-1}) = \tfrac{(x_i+t'_{k-1})}{n-k} p'_{-i}(t'_{k-1}) + p_{-i}(t'_{k-1}) \leq 0$$
$$p(t_k) = \tfrac{(x_i+t_k)}{n-k} p'_{-i}(t_k) + p_{-i}(t_k) \geq 0$$

so by continuity, $p(t)$ must have a root between t'_{k-1} and t_k. Since $0 \leq t'_{k-1}$, this root must be positive, hence $x \notin K_{E_k(\cdot)}$ (see Figure 2.1).

Case 2. Suppose $t'_{k-1} < x_i \leq t_k$. Then we can write

$$p(-x_i) = \tfrac{(x_i+(-x_i))}{n-k} p'_{-i}(-x_i) + p_{-i}(-x_i) \leq 0$$
$$p(t_k) = \tfrac{(x_i+t_k)}{n-k} p'_{-i}(t_k) + p_{-i}(t_k) \geq 0$$

and again by continuity, $p(t)$ must have a positive root, so $x \notin K_{E_k(\cdot)}$.

Case 3. Finally, suppose that $t_k < -x_i$. Then

$$p(t_k) = \tfrac{(x_i+t_k)}{n-k} p'_{-i}(t_k) + p_{-i}(t_k) \leq 0$$
$$p(-x_i) = \tfrac{(x_i+(-x_i))}{n-k} p'_{-i}(-x_i) + p_{-i}(-x_i) \geq 0$$

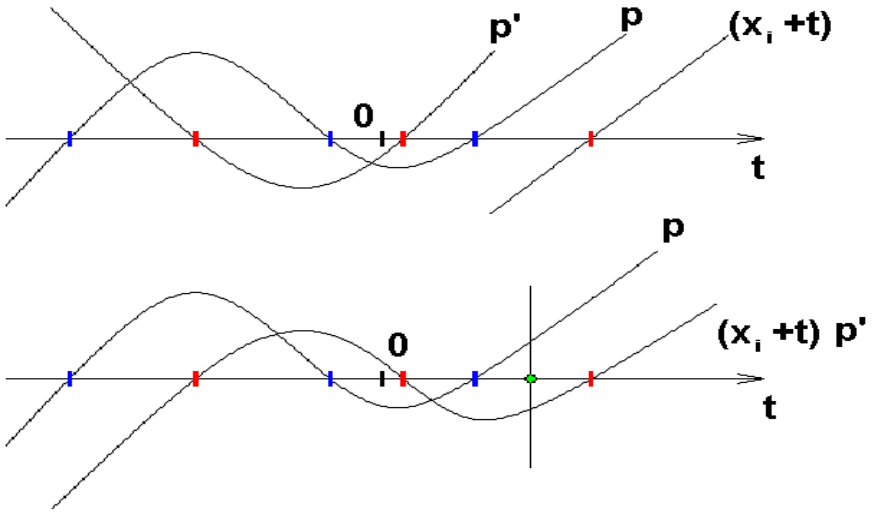

Figure 2.1: Necessary condition for $x \in K_{E_k}$, proof, root interlacing case 1

so by continuity, $p(t)$ must have a positive root and therefore $x \notin K_{E_k(\cdot)}$. □

Corollary 2.5.2 (Necessary conditions for $x \in K_{E_k}$)**.** *Assume* $2 \leq k \leq n$ *and* $x \in K_{E_k(\cdot)}$. *If* $x_i \leq 0$ *then* $x_{-i} \in K_{E_k(\cdot_{-i})}$. *Moreover, if* $x \in \partial K_{E_k(\cdot)}$, $x \notin \mathbb{R}^n_+$, *and* $x_i > 0$ *then* $x_{-i} \notin K_{E_k(\cdot_{-i})}$.

Proof. We write $p(t) = E_k(x + t\mathbf{1}) = \frac{(x_i+t)}{n-k} p'_{-i}(t) + p_{-i}(t)$ and at $t = 0$ we have $E_k(x) = p(0) = \frac{x_i}{n-k} p'_{-i}(0) + p_{-i}(0) = x_i E_{k-1}(x_{-i}) + E_k(x_{-i})$. Since $x_{-i} \in K_{E_{k-1}(\cdot_{-i})}$ by Theorem 2.5.1, we have $p'_{-i}(0) = (n-k)E_k(x_{-i}) \geq 0$ and also from Theorem 2.2.3 $p(0) = E_k(x) \geq 0$. We rearrange terms: $p'_{-i}(0)x_i = (n-k)(E_k(x) - p_{-i}(0))$.

If $x_i \leq 0$ we have $E_k(x) - p_{-i}(0) \leq 0$, so $p_{-i}(0) = E_k(x_{-i}) \geq 0$ and combined with $x_{-i} \in K_{E_{k-1}(\cdot_{-i})}$, this gives us $x_{-i} \in K_{E_k(\cdot_{-i})}$.

Now let $x \in \partial K_{E_k(\cdot)}$, so that $E_k(x) = 0$, and $x_i > 0$. We have two possibilities here. If $p'_{-i}(0) > 0$, then $-p_{-i}(0) > 0$ and hence $x_{-i} \notin K_{E_k(\cdot_{-i})}$. Alternatively, if $p'_{-i}(0) = 0$ ($x \in \partial K_{E_{k-1}(\cdot_{-i})}$), then $p_{-i}(0) = 0$ and $x_{-i} \in \partial K_{E_k(\cdot_{-i})}$, so by

Proposition 2.2.6 $x \in K_{E_{n-1}(\cdot_{-i})} \equiv \mathbb{R}^{n-1}_+$. But we assumed $x \notin \mathbb{R}^n_+$, so this cannot happen. □

To summarize: for $x \in K_{E_k(\cdot)}$

$$x_{-i} \in K_{E_{k-1}(\cdot_{-i})}, \forall i$$

$$\text{if } x_i \leq 0 \text{ then } x_{-i} \in K_{E_k(\cdot_{-i})} \subset K_{E_{k-1}(\cdot_{-i})}$$

$$\text{if } x \in \partial K_{E_k(\cdot)} \setminus \mathbb{R}^n_+ \text{ and } x_i > 0 \text{ then } x_{-i} \in K_{E_{k-1}(\cdot_{-i})} \setminus K_{E_k(\cdot_{-i})}$$

2.5.2 Alternative characterization of the hyperbolicity cones associated with elementary symmetric polynomials

From what we have established it is easy to derive necessary and sufficient conditions for $x \in \mathbb{R}^n_\downarrow$ to be in $K_{E_k(\cdot)}$.

Instead of considering the whole space $\mathbb{R}^n$ we will confine ourselves to the cone $\mathbb{R}^n_\downarrow = \{x \in \mathbb{R}^n : x_n \leq x_{n-1} \leq x_{n-2} \leq \cdots \leq x_1\}$.

Theorem 2.5.3 (Necessary and sufficient conditions for $x \in K_{E_k}$)**.** *Assume $2 \leq k \leq n$ and $x \in \mathbb{R}^n_\downarrow$. Then $x \in K_{E_k(\cdot)}$ iff $x_{-n} \in K_{E_{k-1}(\cdot_{-n})}$ and $E_k(x) \geq 0$.*

Proof. The conditions are necessary (see previous lemma and Theorem 2.2.3). We need to show sufficiency. The case $k = n$ is trivial, so assume $k < n$. If $x_n \geq 0$, then obviously $x \in K_{E_k(\cdot)} (\supset \mathbb{R}^n_+)$, so assume $x_n < 0$.

Let $p(t)$, $p_{-n}(t)$ and $p'_{-n}(t)$ with corresponding roots (including multiplicities) of $p_{-n}(t)$, $\{t_i : i = 1, \ldots, k\}$, and roots of $p'_{-n}(t)$, $\{t'_i : i = 1, \ldots, (k-1)\}$, in non-decreasing order as before. We write $p(t) = E_k(x + t\mathbf{1}) = \frac{(x_n+t)}{n-k} p'_{-n}(t) + p_{-n}(t))$.

Observe

$$p(t_k) = \frac{(x_n + t_k)}{n-k} p'_{-n}(t_k) \leq 0$$

since $(x_n + t_k) \leq 0$ (by the root interlacing $t_k \leq -x_n$) and $p'_{-n}(t_k) \geq 0$ (recall that $p'_{-n}(t) \uparrow \infty$ as $t \uparrow \infty$). Also, $p(0) = E_k(x) \geq 0$ by the assumption. Thus the interval $[t_k, 0]$ must contain at least one root of $p(t)$.

By counting the remaining roots of $p(t)$, $t \leq t_k$ (by looking at sign patterns at the endpoints of intervals $[t_i, t'_i], i = 1, \ldots, (k-1)$) we conclude that $[t_k, 0]$ must contain only one (rightmost) root of $p(t)$ (so there could be no other roots to the right of 0) and hence $x \in K_{E_k(\cdot)}$. □

Corollary 2.5.4 (Necessary and sufficient conditions for $x \in K_{E_k}$)**.** *Assume* $2 \leq k \leq (n-1)$ *and* $x \in \mathbb{R}^n_\downarrow$. *Then* $x \in K_{E_k(\cdot)}$ *iff* $x_{-n} \in K_{E_k(\cdot_{-n})}$ *and* $E_k(x) \geq 0$.

Proof. Straightforward. □

Note that this implies (for $2 \leq k \leq (n-1)$)

(i) $x \in K_{E_k}$,

(ii) $x_{-n} \in K_{E_{k-1}(\cdot_{-n})}$ and $E_k(x) \geq 0$,

(iii) $x_{-n} \in K_{E_k(\cdot_{-n})}$ and $E_k(x) \geq 0$,

are all equivalent.

Example: consider $E_2(x)$ in $\mathbb{R}^2$: $x_1x_2 + x_1x_3 + x_2x_3$. Let us compare (ii) and (iii): (ii) gives us

$$x_1x_2 + x_1x_3 + x_2x_3 \geq 0$$

$$x_3 \leq 0, \;\; x_1 + x_2 \geq 0$$

The first constraint can be rewritten as $(x_1 + x_2)x_3 + x_1x_2 \geq 0$, implying now x_1 and x_2 must be of the same sign. Combined with $x_1 + x_2 \geq 0$, this gives us $x_1, x_2 \geq 0$, that is (iii). The reverse implication is trivial.

We make one observation about $\mathbb{R}^n_\downarrow$, namely, for an arbitrary index i and any $k \geq 0$ we have $x_{-i} \leq x_{-(i+k)}$ (easy to check).

Let $q(\cdot_{-i}) = \frac{E_k(\cdot_{-i})}{E_{k-1}(\cdot_{-i})} = (n-k)\frac{p_{-i}(0)}{p'_{-i}(0)}$ and recall that this function was shown to be concave on $K_{E_{k-1}(\cdot_{-i})}$ and is homogeneous of degree one (i.e., for $t \in \mathbb{R}$ we have $q(tx) = tq(x)$ since $E_k(x)$ is k-homogeneous and $E_{k-1}(x)$ is $(k-1)$-homogeneous).

Proposition 2.5.5. *Assume* $2 \leq k \leq n$. *Let* $q(\cdot) = \frac{E_k(\cdot)}{E_{k-1}(\cdot)}$. *If* $x, y \in K_{E_{k-1}(\cdot)}$, *then* $q(x+y) \geq q(x) + q(y)$.

Proof. Since $q(x)$ is concave (see Theorem 2.4.3) we can write $-q(\frac{x+y}{2}) \leq \frac{-q(x)-q(y)}{2}$ and from homogeneity it follows that $-q(x+y) \leq -q(x) - q(y)$. □

Remark 2.5.6 (On a set of necessary conditions, compare with Corollary 2.5.2). Assume $2 \leq k \leq (n-1)$ and $x \in \mathbb{R}^n_\downarrow$. If $x \in \partial K_{E_k(\cdot)}$, then $\exists j$ s.t.

$$x_{-i} \in K_{E_k(\cdot_{-i})} \quad \text{for} \quad i \geq j$$
$$x_{-i} \notin K_{E_k(\cdot_{-i})} \quad \text{for} \quad i < j$$

Proof. For fixed i and $k \geq 0$, $x_{-i} \leq x_{-(i+k)}$. Observe $q(x_{-(i+k)}) = q(x_{-i} + (x_{-(i+k)} - x_{-i})) \geq q(x_{-i}) + q(x_{-(i+k)} - x_{-i}) \geq q(x_{-i})$ since $(x_{-(i+k)} - x_{-i}) \in \mathbb{R}^n_+$. The condition on x_{-i} being in or out of $K_{E_k(\cdot_{-i})}$ for $i < n$ (i.e., $q(\cdot_{-i})$ having the right sign) is implied by "monotonicity" of $q(\cdot_{-i})$. □

2.5.3 First derivative cone for $\mathbb{R}^n_+$ and its dual

Recall for any hyperbolic polynomial h (w.r.t. d, w.l.o.g. $h(d) > 0$, of degree m) one can give a characterization of the (closure of) associated hyperbolicity cone

K_h by the set of polynomial inequalities of the form:

$$\begin{cases} h(x) \geq 0 \\ h'(x) \geq 0 \\ \vdots \\ h^{(m-1)}(x) \geq 0 \end{cases}$$

For $p(x) := E_n(x)$ and its derivative $p'(x) := E_{n-1}(x)$ we claim that $x \in K_{p'}$ iff $p'(x) \geq 0$ and at most one $x_i < 0$ with the rest $x_j \geq 0$ for $i \neq j$. This characterization follows from our necessary and sufficient conditions (see Corollary 2.5.4):

E_{n-1}	E_k
$x_n \leq x_{n-1} \leq x_{n-2} \leq \cdots \leq x_1$	$x_n \leq x_{n-1} \leq x_{n-2} \leq \cdots \leq x_1$
Necessary and sufficient conditons: $\begin{cases} x_{-n} \in K_{E_{n-1}(\cdot_{-n})} \Leftrightarrow x_i \geq 0 \text{ for } i < n \\ E_{n-1}(x) \geq 0 \end{cases}$	Necessary and sufficient conditions: $\begin{cases} x_{-n} \in K_{E_k(\cdot_{-n})} \\ E_k(x) \geq 0 \end{cases}$

We are going to construct a representation of the dual cone to $K_{p'} \equiv K_{E_{n-1}}$ using this characterization.

In part, we will also rely on the following observation.

Proposition 2.5.7. *If $K \subseteq \mathbb{R}^n$ is a cone admitting a decomposition into (smaller) cones $\{K_i\}_{i \in I}$, $K = \bigcup_{i \in I} K_i$, then its dual cone satisfies $K^* = \bigcap_{i \in I} {K_i}^*$.*

Proof. Straightforward from the definition of the dual cone. □

We form a (disjoint-interior) partitioning for $K_{p'}$ in the following manner: $K_{p'} = (\bigcup_{i=1\ldots n} K^i_{p'}) \bigcup K^0_{p'}$ where $K^i_{p'} = \{x \in \mathbb{R}^n : x_i \leq 0, x_j \geq 0, j \neq i, p'(x) \geq 0\}$ and $K^0_{p'} = K_p = (K_p)^* = \mathbb{R}^n_+$, claiming that each of the $K^i_{p'}$ admits SDR representation (with strictly-feasible solution), see Figure 2.2. Based on Corollary 2.3.5 it is now

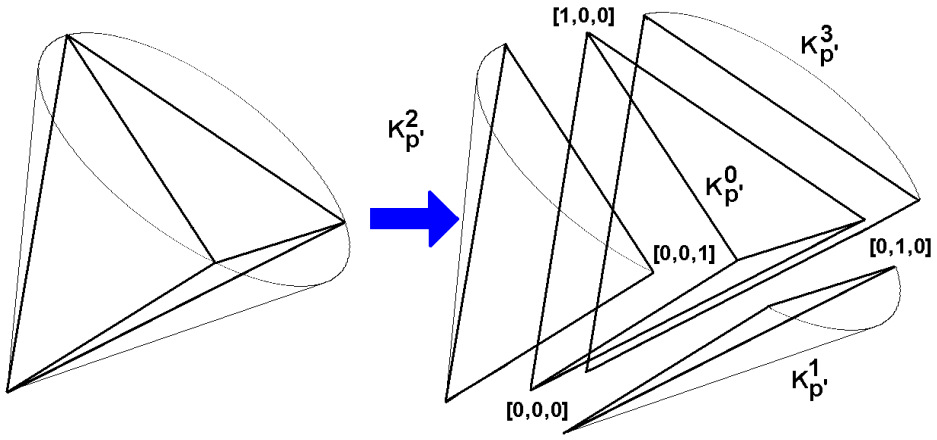

Figure 2.2: $K_{E_{n-1}}$ cone decomposition in $\mathbb{R}^3$

easy to reconstruct the dual cone (as an affine image/section of the positive semi-definite cone).

It is left to demonstrate how to represent each $K^i_{p'}$ via linear matrix inequality (LMI). We show how to do this for $K^1_{p'}$.

Consider the matrix $W_1(x)$ of the form:

$$\begin{bmatrix} -x1 & -x1 & -x1 & \cdots & -x1 \\ -x1 & x2 & 0 & \cdots & 0 \\ -x1 & 0 & x3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -x1 & 0 & 0 & \cdots & x_n \end{bmatrix}$$

The condition of the form $W_1(x) \succeq 0$ is clearly an LMI. Recall that for a real symmetric matrix to be positive semi-definite it is necessary and sufficient that all its principal minors have nonnegative determinants. Proceed by evaluating these determinants from the bottom-right corner. We get that all $x_j, j = 2, \ldots, n$ must be ≥ 0 and the last determinant (of $W_1(x)$ itself) being nonnegative is equivalent

to $-x_1 p'(x) \geq 0$, which, combined with $-x_1 \geq 0$ (consider the fist principal minor of $W_1(x)$, $\det[-x_1] \geq 0$), implies $p'(x) \geq 0$. In order to convince ourselves that $\det(W_1(x)) = -x_1 p'(x)$, evaluate this determinant using algebraic complements of the first row. Denote $W_1^{ij}(x)$ the matrix obtained from $W_1(x)$ by omitting its i^{th} row and j^{th} column. We can write $\det(W_1(x)) = \sum_{j=1}^{n}(-1)^{j+1}\det(W_1^{1j}(x))$ so that $\det(W_1^{11}(x)) = \prod_{j=2}^{n} x_j$ and $\det(W_1^{1j}(x)) = (-1)^{j+1}\prod_{k\neq j} x_k$ observing that any $W_1^{1j}(x)$ for $j \geq 3$ can be put in a lower-triangular form by making $(j-2)$ row permutations corresponding to switching k^{th} and $(k+1)^{th}$ column with each other sequentially, bottom to top, starting with $k = j-2$, which will result in $(j-2)$ sign changes of the corresponding determinant. Clearly, strict feasibility for this LMI is insured as well (e.g., take $x_2 = x_3 = \cdots = x_n = 1, x_1 < 0$, with $|x_1|$ small enough). So Corollary 2.3.5 can be applied to get $(K_{p'}^1)^*$ as an SDR set.

Finally, to get the representation of the dual cone to $K_{E_{n-1}(\cdot)}$ take the intersection of the dual cones corresponding to its components: $(K_{E_{n-1}(\cdot)})^* = (\cap_{i=1,\ldots,n}\{x \in \mathbb{R}^n : W_i(x) \succeq 0\}^*) \cap (\mathbb{R}^n_+)^*$.

To illustrate this idea we will consider the derivation of $(K_{E_{n-1}(\cdot)})^*$ in $\mathbb{R}^3$, which is perhaps not the most exciting example (it is just a quadratic cone after all) but is quite an illustrative one (it is easy to appeal to geometric interpretation of the results).

The dual cone can be given by $(\cap_{i=1,\ldots,n}\{x \in \mathbb{R}^n : W_i(x) \succeq 0\}^*) \cap (\mathbb{R}^n_+)^*$. Consider $\{x \in \mathbb{R}^n : W_1(x) \succeq 0\}$ first:

$$W_1(x) : \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \mapsto \begin{bmatrix} -1 & -1 & -1 \\ -1 & & \\ -1 & & \end{bmatrix} x_1 + \begin{bmatrix} 0 & & \\ & 1 & \\ & & 0 \end{bmatrix} x_2 + \begin{bmatrix} 0 & & \\ & 0 & \\ & & 1 \end{bmatrix} x_3$$
$$= A_1 x_1 + A_2 x_2 + A_3 x_3$$

(think of as $(\sum_{i=1}^{n} x_i A_i + \sum_{j=1}^{m} u_j B_j) + B$ with B_j, B being 0).

This gives the following representation of $\{x \in \mathbb{R}^n : W_i(x) \succeq 0\}^*$:

$$\langle \begin{bmatrix} -1 & -1 & -1 \\ -1 & & \\ -1 & & \end{bmatrix}, \Lambda_1 \rangle = y_1, \quad \langle \begin{bmatrix} 0 & & \\ & 1 & \\ & & 0 \end{bmatrix}, \Lambda_1 \rangle = y_2, \quad \langle \begin{bmatrix} 0 & & \\ & 0 & \\ & & 1 \end{bmatrix}, \Lambda_1 \rangle = y_3,$$
$$\Lambda_1 \succeq 0$$

and similarly we can derive the expressions for $\{x \in \mathbb{R}^n : W_2(x) \succeq 0\}^*$ (with Λ_2) and $\{x \in \mathbb{R}^n : W_3(x) \succeq 0\}^*$ (with Λ_3). At this point we can reconstruct the dual cone to $K_{E_2(\cdot)}$ as a collection of three sets of LMI's each corresponding to $\{x \in \mathbb{R}^n : W_i(x) \succeq 0\}^*$ (with same y in all of them, $i = 1, 2, 3$). Note that there is no need to further restrict ourselves to $y \in \mathbb{R}^3_+$ since this is already implied by the constraints.

An interesting question that remains unanswered is, "How would one get the (complete) representation of the original cone $K_{E_2(\cdot)}$ in terms of LMI's (recall that we had represented only parts of it so far)?". To do this we take the dual of $K^*_{E_{n-1}(\cdot)}$. Firstly, let us switch from the image of a positive semi-definite cone to its affine slice in each of the $\{x \in \mathbb{R}^n : W_i(x) \succeq 0\}^*$. Starting with $\{x \in \mathbb{R}^n : W_1(x) \succeq 0\}^*$, fixing a basis in $S^{3\times 3}$ to be

$$\{B_i\}_{i=1}^{6} = \left\{ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \right.$$
$$\left. \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right\}$$

we can rewrite (substituting $\Lambda_1 = \sum_{j=1}^{6} B_j \lambda_{1j}$)

$$\langle A_i, \Lambda_1 \rangle = y_i \Leftrightarrow \langle A_i, \sum_{j=1}^{6} B_j \lambda_{1j} \rangle = y_i \Leftrightarrow \sum_{j=1}^{6} \langle B_j, A_i \rangle \lambda_{1j} = y_i$$

so we get the following system of constraints

$$\sum_{j=1}^{6} \langle B_j, A_1 \rangle \lambda_{1j} = (-1)\lambda_{11} + (-2)\lambda_{12} + (-2)\lambda_{13} + (0)\lambda_{14} + (0)\lambda_{15} + (0)\lambda_{16} = y_1$$

$$(1)\lambda_{14} = y_2$$

$$(1)\lambda_{16} = y_3$$

$$\Lambda_1 \succeq 0$$

This can be written as

$$\begin{bmatrix} -y_1 - 2(\lambda_{12} + \lambda_{13}) & \lambda_{12} & \lambda_{13} \\ \lambda_{12} & y_2 & \lambda_{15} \\ \lambda_{13} & \lambda_{15} & y_3 \end{bmatrix} \succeq 0$$

Similarly, we can transform two other LMI's (corresponding to $W_2(x)$ and $W_3(x)$). The complete description of the dual cone is

$$\begin{bmatrix} -y_1 - 2(\lambda_{12} + \lambda_{13}) & \lambda_{12} & \lambda_{13} \\ \lambda_{12} & y_2 & \lambda_{15} \\ \lambda_{13} & \lambda_{15} & y_3 \end{bmatrix} \succeq 0, \begin{bmatrix} -y_1 - 2(\lambda_{22} + \lambda_{23}) & \lambda_{22} & \lambda_{23} \\ \lambda_{22} & y_2 & \lambda_{25} \\ \lambda_{23} & \lambda_{25} & y_3 \end{bmatrix} \succeq 0$$

$$\begin{bmatrix} -y_1 - 2(\lambda_{32} + \lambda_{33}) & \lambda_{32} & \lambda_{33} \\ \lambda_{32} & y_2 & \lambda_{35} \\ \lambda_{33} & \lambda_{35} & y_3 \end{bmatrix} \succeq 0$$

(we can think of it as one matrix $\succeq 0$ with the off-diagonal block entries being zeros and the three diagonal blocks as specified above).

Now we can apply the same procedure to take the dual of the dual cone (to get the primal cone itself; note again that the constraint is strictly feasible, for example,

take $y = \mathbf{1}, -1/2 < \lambda_{i,j} < -1/3, \forall i, j)$. By rewriting the constraints corresponding to the dual cone in the form $\sum_{i=1}^{3} y_i \widetilde{A}_i + \sum_{j=1}^{3} (\lambda_{j2}\widetilde{B}_{j2} + \lambda_{j3}\widetilde{B}_{j3} + \lambda_{j5}\widetilde{B}_{j5}) \succeq 0$ we get

$$\begin{array}{|ccc|ccc|ccc|}
\hline
\mu_{1,11} & \mu_{1,11} & \mu_{1,11} & & & & & & \\
\mu_{1,11} & \mu_{1,22} & & & & & & & \\
\mu_{1,11} & & \mu_{1,33} & & & & & & \\
\hline
& & & \mu_{2,11} & \mu_{2,11} & \mu_{2,11} & & & \\
& & & \mu_{2,11} & \mu_{2,22} & & & & \\
& & & \mu_{2,11} & & \mu_{2,33} & & & \\
\hline
& & & & & & \mu_{3,11} & \mu_{3,11} & \mu_{3,11} \\
& & & & & & \mu_{3,11} & \mu_{3,22} & \\
& & & & & & \mu_{3,11} & & \mu_{3,33} \\
\hline
\end{array} \succeq 0$$

$$x_1 = \mu_{1,33} + \mu_{2,33} - \mu_{3,11}$$

$$x_2 = \mu_{1,22} - \mu_{2,11} + \mu_{3,22}$$

$$x_3 = -\mu_{1,11} + \mu_{2,22} + \mu_{3,33}$$

for $x \in K_{E_2}$, where the off-diagonal blocks are not necessarily zeroes anymore. At this point we make a few observations. Firstly, note that for this matrix to be positive semi-definite it is enough to consider only the three specified block-diagonal entries. Namely, it is enough to require these three diagonal blocks to be positive semi-definite (follows from the positive semi-definiteness criteria using minors) while setting the off-diagonal blocks to 0. So this constraint is "decomposable" into three independent LMI's (corresponding to these three blocks) which are further "assembled" together in order to get the primal variables x_1, x_2, x_3. Secondly, there is a very simple interpretation to this set of constraints. Observe

that each of the blocks ($i = 1, 2, 3$)

$$\begin{bmatrix} \mu_{i,11} & \mu_{i,11} & \mu_{i,11} \\ \mu_{i,11} & \mu_{i,22} & \\ \mu_{i,11} & & \mu_{i,33} \end{bmatrix} \succeq 0$$

corresponds to $K^i_{p'} = \{x \in \mathbb{R}^n : x_i \leq 0, x_j \geq 0, j \neq i, E_{n-1}(x) \geq 0\} = \{x \in \mathbb{R}^n : W_i(x) \succeq 0\}$ but with x's now renamed into $\pm\mu$'s. Therefore, each block describes just one of these "slabs".

The remaining linear constraints

$$x_1 = \mu_{1,33} + \mu_{2,33} - \mu_{3,11}$$

$$x_2 = \mu_{1,22} - \mu_{2,11} + \mu_{3,22}$$

$$x_3 = -\mu_{1,11} + \mu_{2,22} + \mu_{3,33}$$

are building a (convex) combination of these slabs ($K_{E_{n-1}(\cdot)}$ is a cone so we can assume that the points have unit weight). To conclude, this set of constraints simply tells us that we can obtain any point in the cone itself as a convex combination of the points in $K^i_{p'} = \{x \in \mathbb{R}^n : x_i \leq 0, x_j \geq 0, j \neq i, E_{n-1}(x) \geq 0\} = \{x \in \mathbb{R}^n : W_i(x) \succeq 0\}$, $i = 1, \ldots, n$ (see Figure 2.3). Once we had the description of the slabs, we could have immediately written out the description for the whole $K_{E_{n-1}(\cdot)}$.

Remark 2.5.8 (Concluding comments). We have demonstrated how one can obtain a SDR representation for the cone in $\mathbb{R}^n$ corresponding to E_{n-1}, i.e., what we refer to as the first derivative cone to the nonnegative orthant. Furthermore, using this semi-definite embedding we demonstrated how one can easily characterize the corresponding dual cone $K^*_{E_{n-1}}$, which was previously unknown. It should be noted that since the dual cone $K^*_{E_{n-1}}$ was constructed as an affine section of S^k_+,

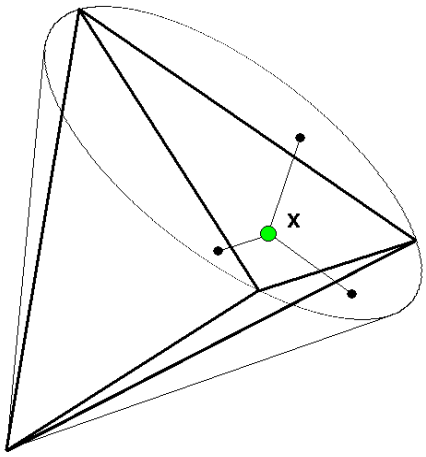

Figure 2.3: $K_{E_{n-1}}$ (primal) SDR in $\mathbb{R}^3$

this approach provides us with a way to get a self-concordant barrier for the dual cone as well.

Using somewhat similar techniques I obtained a partial description of a cone $K_{E_{n-2}}$, but since this description is incomplete as of yet and is work in progress it is not presented here.

Chapter 3

Shrink-Wrapping algorithm for linear programming

In this chapter we present a new framework for linear programming problems recently introduced by James Renegar, which relies on consecutive relaxations of the underlying hyperbolicity cone for linear programming ($\mathbb{R}^n_+ \equiv K_{E_n,\mathbf{1}}$). The algorithmic approach can be viewed as a generalization of IPM.

Although most of the ideas presented stay valid for more general hyperbolic programming problems, we will focus on the case of linear programming for the sake of concreteness.

We will present the general framework followed by analysis of the continuous setting establishing some important properties. The analysis will culminate with the establishment of local convergence properties (of one particular variant) of the resulting discrete algorithm, namely, showing that it converges (R-)super-linearly (under appropriate assumptions).

3.1 The general framework and convergence

Let $c \in \mathbb{R}^n, A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$ and consider

$$(P) \quad \begin{cases} \min c^T x \\ Ax = b \\ x \in \mathbb{R}^n_+ \end{cases}$$

We assume that (P) is bounded and has a unique optimal solution x^*. Suppose further that we have a point d which is strictly feasible for (P) $(Ad = b, d > 0)$.

Pick $1 \leq k \leq n$. Consider the k^{th} elementary symmetric function in $\mathbb{R}^n$, E_k, evaluated at a point x/d, $x \mapsto E_k(x/d)$. Recall that given such (fixed) d, $E_k(x/d)$ will correspond to the $(n-k)^{th}$ derivative hyperbolic polynomial of E_n with respect to direction d (up to a multiplicative constant), that is, $E_n^{(n-k)}(x)$ ($\equiv E_n^{(n-k)}(d,x)$).

Associated with $E_n^{(n-k)}(x)$ (and d) we have (the closure of) a hyperbolicity cone $K_{E_n^{(n-k)},d}$ with $\mathbb{R}^n_+ \subseteq K_{E_n^{(n-k)},d} \equiv K_{k,d}$ (see Corollary 2.2.4).

Consider the optimization problem

$$(P(d)) \quad \begin{cases} \min c^T x \\ Ax = b \\ x \in K_{k,d} \end{cases}$$

a relaxation of (P). For any $d \in \mathbb{R}^n_{++}$, this is a well-defined convex problem. We denote its optimal solution as $x(d)$ (if it exists).

Fact 3.1.1 (Renegar)**.** *$x(d)$ is unique provided $x(d) \notin \mathbb{R}^n_+$.*

Proof. Follows from the boundary of the corresponding cone $K_{k,d}$ having strict curvature unless it coincides with some boundary face of $\mathbb{R}^n_+$ (see [8], Theorem 14; note $\mathbb{R}^n_+$ is regular). □

Remark 3.1.2. For the time being we will be assuming that it is "easy" to obtain a solution $x(d)$ for $(P(d))$. Our goal is to find x^* – the solution for (P). Bearing this in mind, note that in the proposition above the uniqueness of $x(d)$ is guaranteed across all of the domain of interest, namely if $x(d)$ is already in $\mathbb{R}^n_+$ then $x(d)$ is also the solution for (P), $x(d) = x^*$.

To get a better idea of how (P) and $(P(d))$ relate to each other let us consider one example. Consider solving the following LP

$$\{\min c^T x : [1;1;1]^T x = 3, x \in \mathbb{R}^3_+\}$$

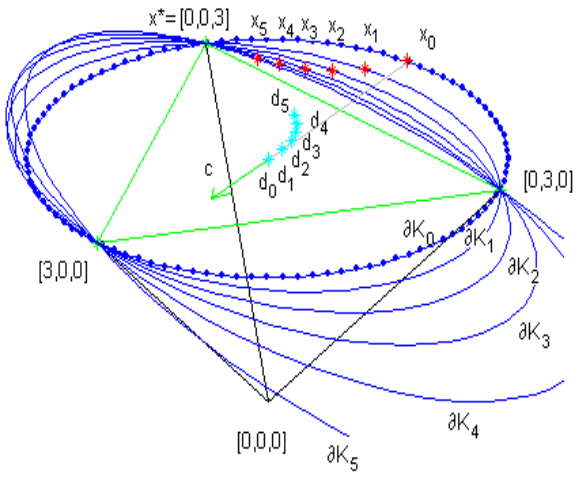

Figure 3.1: Shrink-Wrapping algorithm for LP, relating (P) and $(P(d))$

with some $c \in \mathbb{R}^3$ (see Figure 3.1).

For the corresponding $(P(d))$ we take $K_{k,d}$ to be the first derivative cone of the nonnegative orthant (i.e. the quadratic cone) w.r.t. direction d, that is $K_{2,d}$. Note that in this example the resulting cone touches on the axes of the nonnegative orthant.

Given $d_0 = \mathbf{1}$, compute the solution $x_0 = x(d_0) \in \partial K_0$ to $(P(d_0))$ (∂K_0 is the boundary of the cone corresponding to d_0; we adapt a slightly shorter notation here, denoting K_{2,d_0} as just K_0; we depict the boundary intersected with the affine constraint in Figure 3.1). Obviously, x_0 and the true LP solution $x^* = (0, 0, 3)$ are quite far from one another. How can we bring the solution of $(P(d))$ closer to the solution of (P)?

Let us pick a new hyperbolicity direction d_1 to be on the line between the current hyperbolicity direction d_0 and the corresponding optimal solution x_0, shifted slightly from the point d_0 towards x_0. Corresponding to this new direction d_1, we

have a (slightly) different cone K_1 with the corresponding (intersection of the affine constraints with its) boundary ∂K_1, and a new optimal solution $x(d_1) = x_1 \in \partial K_1$. Note that now the solution for the new relaxation $(P(d_1))$ is closer to x^* than that of $(P(d_0))$.

Having d_1 and x_1, now pick d_2 in a similar fashion (to be on the line between these two points slightly shifted from d_1 towards x_1) to obtain $x(d_2) = x_2$, and so on. Continuing in this manner, both $\{d_i\}_{i\geq 0}$ and $\{x(d_i)\}_{i\geq 0}$ will tend to converge to the optimal LP solution x^*. It happens that the convergence is not particular to this example.

The dynamics can be formalized through the ODE

$$\begin{aligned} \dot{d}(t) &= x(d(t)) - d(t) \\ d(0) &= d \end{aligned} \tag{3.1.0.1}$$

Assume that a starting point d is chosen such that there exists $x(d)$ $((P(d))$. For brevity, denote $x(d(t))$ as just $x(t)$.

Theorem 3.1.3 (Renegar)**.** *Under dual strict feasibility for* (P)

$$\lim_{t\uparrow\infty} d(t) = x^*$$

Conjecture: *In addition,* $\lim_{t\uparrow\infty} x(t) = x^*$

To gain a sense why this would be a reasonable choice for the dynamics of $d(t)$, simply observe

$$c^T\dot{d}(t) = c^T x(d(t)) - c^T d(t) < 0$$

since $d(t)$ is strictly feasible for the relaxed problem of which $x(d(t))$ is optimal.

This setting has a connection with interior-point methods. To explain, we introduce the notion of a *central swath*

$$\mathrm{CS}_k(P) := \{d \in \mathbb{R}^n_{++} : (P(d)) \text{ corresp. to } K_{k,d} \text{ has an optimal solution}\}$$

The following proposition can be made:

Proposition 3.1.4 (Renegar)**.** *For $k = 1$, $CS_k(P)$ is the central path for (P) (corresponding to $-\ln \prod_{i=1}^{n} x_i$).*

Proposition 3.1.5 (Renegar)**.** *If $k = 2$ and $d(0)$ is on the central path for (P), then $\{d(t), t \geq 0\}$ is (part of) the central path for (P).*

Remark 3.1.6. Unlike interior-point methods, we explicitly trace $\{x(d(t)), t \geq 0\}$. Moreover, in the example above (see Figure 3.1), in our experience more generally, and as the choice of the dynamics for $d(t)$ might suggest, $x(t) = x(d(t))$ seems to identify the optimal solution x^* "sooner" than $d(t)$ itself (the direction of change for $d(t)$ is in a way "pointed to" by $x(t)$). This behavior is even more striking in the case of LP when $\{d(t), t \geq 0\}$ coincides with the central path making a sharp turn towards x^* while going almost orthogonal to the face of $\mathbb{R}^n_+$ containing x^* first; see Figure 3.2.

This observation suggests that we should concentrate on studying the properties of $\{x(t), t \geq 0\}$ rather than of $\{d(t), t \geq 0\}$.

The proposed scheme for the corresponding discrete algorithm is to:

follow the "path" iteratively, generating a sequence of pairs, (d_i, x_i), $i = 1, \ldots, \infty$:

1) given direction d_i, compute (approximate) optimal solution $x_i \approx x(d_i)$ for the relaxation,

2) determine d_{i+1} by making a step from d_i towards x_i,

iterate.

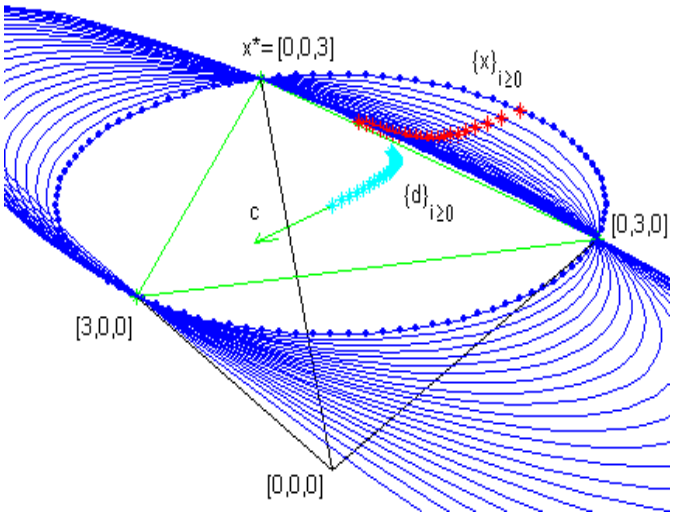

Figure 3.2: Shrink-Wrapping algorithm for LP, $d(t)$ versus $x(t)$

Remark 3.1.7 (Note on the change of coordinates for $\{x : Ax = b\}$). Let B be a surjective linear map $B : \mathbb{R}^l \to \text{null}(A)$ and assume $x_0 \in \mathbb{R}^n$ satisfies $Ax_0 = b$. Thus, any affine feasible point x can be written as $x = x_0 + B\widetilde{x}$ (for $\widetilde{x} \in \mathbb{R}^l$). The ODE 3.1.0.1 becomes

$$B\dot{\widetilde{d}}(t) = B\widetilde{x}(\widetilde{d}(t)) - B\widetilde{d}(t) + (x_0 - x_0)$$
$$B\widetilde{d}(0) = d - x_0$$

(with $d(t) = x_0 + B\widetilde{d}(t)$ and $x(\widetilde{d}(t)) = x_0 + B\widetilde{x}(\widetilde{d}(t))$ being a corresponding solution of $(P(d))$). Moreover, if B is injective, this is equivalent to

$$\dot{\widetilde{d}}(t) = \widetilde{x}(\widetilde{d}(t)) - \widetilde{d}(t)$$
$$\widetilde{d}(0) = \widetilde{d}$$

As a consequence, we can pick an arbitrary affine coordinate system for $\{x : Ax = b\}$ and analyze the behavior of 3.1.0.1 in this particular coordinate system. We will refer to this as "affine invariance" of 3.1.0.1.

3.2 The precise setting: main properties

3.2.1 More observations and the setting

So far, all the statements we have made about the new framework for LP can be translated almost directly to the case of more general hyperbolic programming problems

$$(P) \quad \begin{cases} \min\langle c, x\rangle \\ Ax = b \\ x \in K_{p,d} \end{cases}$$

($c \in \mathbb{R}^n, b \in \mathbb{R}^m$, a linear map $A : \mathbb{R}^n \to \mathbb{R}^m$ and p – a hyperbolic polynomial of degree m on $\mathbb{R}^n$ with respect to some $d \in \mathbb{R}^n$). All we need to do is replace the derivative cone of the nonnegative orthant in $(P(d))$ by the corresponding r^{th} derivative cone (say, $r = m - k$ for some k, to keep the notation consistent with LP) of $K_{p,d}$, that is, by $K_{p^{(r)},d}$

$$(P(d)) \quad \begin{cases} \min\langle c, x\rangle \\ Ax = b \\ x \in K_{p^{(r)},d} \end{cases}$$

Before we introduce the refined setting for LP that is the subject of our detailed analysis, we answer the following: given a point $x \in \mathbb{R}^n$ how can we recognize that x is actually $x(d)$?

Since (P) is a convex minimization problem, the optimal solution x^* (if it exists) will necessarily lie on the (relative) boundary of the feasible set $\{x \in \mathbb{R}^n : Ax = b\} \bigcap \partial K_{p,d}$, so we must have $p(x^*) = 0$.

As a consequence of Proposition 2.2.6 we can state the following

Corollary 3.2.1. *If (for some $1 \leq r \leq (m-2)$) the solution of $(P(d))$, $x(d)$,*

satisfies $p^{(r+1)}(x(d)) = 0$*, then* $x(d)$ *also solves* (P)*.*

Proof. Since $x(d) \in K_{p,d}$, $x(d)$ is feasible for (P). Since $(P(d))$ is a relaxation of (P) we thus are done. □

Remark 3.2.2. Suppose $x \neq x^*$. Then we can rewrite the conditions for $x \in \partial K_{p^{(r)},d}$ in terms of $q_d(x) = \frac{p^{(r)}}{p^{(r+1)}}(x)$ (concave on $K_{p^{(r+1)},d}$, see Theorem 2.4.3, rational functional discussed before) since the denominator cannot vanish. Namely $q_d(x) = 0, x \in \text{int}K_{p^{(r+1)},d}$, will correspond to $x \in \partial K_{p^{(r)},d}$.

For convex minimization, the KKT conditions are necessary and sufficient, and thus we can characterize the solution of $(P(d))$ as follows (assuming $x \in \text{int}K_{p^{(r+1)},d}$):

$$\begin{cases} \nabla_x q_d(x) - \tau c \in \text{null}(A) \\ q_d(x) = 0 \\ Ax = b \end{cases}$$

(the necessity follows from a constraint qualification being met at $x(d)$; in particular, $x(d)$ is a regular point: w.l.o.g., $p(d) > 0$, the conic constraint $x \in K_{p^{(r)},d}$ can be rewritten as a set of polynomial inequalities each being $p^{(r+i)}(x) \geq 0, 0 \leq i \leq (m - r - 1)$ and by our assumption $x(d) \neq x^*$ so only $p^{(r)}(x) = 0$ – we have only one active non-linear constraint and the gradient of that constraint at x is non-zero).

This characterization also suggests a way to "trace" the path $\{x_i\}_{i\geq 0}$ given $\{d_i\}_{i\geq 0}$: use a Newton's method-like procedure while changing d_i's ever so slightly.

Let us switch to the case of LP now. How do we pick the degree k of the derivative polynomial $E^{(n-k)}(d, x)$ in $(P(d))$?

Fact 3.2.3. $K_{k,d}$ *has boundary faces only of dimensions* $1, 2, \ldots, k-1$*. For* $j =$

$2, \ldots, k-1$, *the faces are precisely the* j*-dimensional faces of the nonnegative orthant.*

Proof. Follows from the strict curvature of the boundary of this cone except for its (flat) $\partial\mathbb{R}^n_+$ part (see [8], Corollary 17).

To see why $\partial K_{k,d}$ will be flat on $\{x \in \mathbb{R}^n_+ : x_{j_1} = 0, x_{j_2} = 0, \ldots, x_{j_{n-(k-i)}} = 0, 0 \leq j_1 < j_2 < \cdots < j_{n-(k-i)} \leq n\}$ for some $i \geq 1$, the $(k-i)-$dimensional boundary face of $\mathbb{R}^n_+$, observe that x on this set has all elements with at least $(n-(k-1))$ coordinates being 0. Evaluating $E^{(n-k)}(x)$ on this set (e.g., elementary symmetric polynomial of degree k at a point x/d) gives 0 (every monomial term will vanish). Since the cone gives a relaxation to $\mathbb{R}^n_+$, this must be a boundary face of the cone as well. □

We will be analyzing the following LP setting: let

$$(P) \quad \begin{cases} \min c^T x \\ Ax = b \\ x \in \mathbb{R}^n_+ \end{cases}$$

where $c \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^n$. Assume (P) is strictly feasible and has a unique optimal solution x^* which is also non-degenerate, i.e., suppose x^* has exactly m positive components (w.l.o.g., $x_1, x_2, \ldots, x_m > 0$). We will not require (P) to have strictly feasible dual.

Fix $k = m + 1$ and consider the relaxation

$$(P(d)) \quad \begin{cases} \min c^T x \\ Ax = b \\ x \in K_{k,d} \end{cases}$$

(for some $d \in \mathbb{R}^n_{++}$ satisfying $Ad = b$). Note that under these assumptions the cone $K_{k,d}$ will be the first cone (as the degree k of the corresponding polynomial

increases) to touch the nonnegative orthant exactly at the optimal solution x^* (i.e., $k = \arg\min\{k \geq 0 : E^{(n-k)}(d, x^*) = 0\}$). Thus, $(P(d))$ will be a relaxation to (P) whose solution $x(d)$ may coincide with the optimal LP solution x^* for some (properly chosen) d. We will call such a polynomial $E^{(n-k)}(d, x)$ the "wrapping polynomial" for (P), denote it $\widehat{p}(d, x)$, or simply $\widehat{p}(x)$ if $d \equiv \mathbf{1}$. Since $\widehat{p}'(d, x^*) \neq 0, \forall d > 0$, (any) solution $x(d)$ to $(P(d))$ can be characterized using $q_d(x)$ (see remark 3.2.2).

Assume $A : \mathbb{R}^n \to \mathbb{R}^m$ is surjective. Denote $L := \text{null}(A)$, and recall that the ODE 3.1.0.1

$$\dot{d}(t) = x(d(t)) - d(t)$$
$$d(0) = d$$

is affine invariant.

We choose the following coordinate system for $\{x : Ax = b\}$:

$$y := [x_{m+1}, x_{m+2}, \ldots, x_n]$$

(last $(n - m)$ coordinates of x). We treat y as "free" variables. Similarly we introduce

$$e := [d_{m+1}, \ldots, d_n]$$

Note that the optimal solution for (P) is now $y^* = 0$. The ODE 3.1.0.1 can be rewritten as:

$$\begin{aligned} \dot{e} &= y(e) - e \\ e(0) &= e_0 \end{aligned} \tag{3.2.1.1}$$

(Moreover, if we assume that (P) has a strictly feasible dual, then as $t \to \infty$ we must have $e(t) \to 0$, see Theorem 3.1.3.)

We can rewrite the affine feasibility constraints $Ax = b$ using $M = \{1, 2, \ldots, m\}$ and $M^c = \{1, 2, \ldots, n\} \setminus M = \{m+1, \ldots, n\}$ as $A_M x_M + A_{M^c} x_{M^c} = b$, or equivalently,

$$x_M = A_M^{-1}(b - A_{M^c} x_{M^c}) = A_M^{-1} b - A_M^{-1} A_{M^c} y = \widetilde{b} + \widetilde{A} y$$

with $\widetilde{b} = A_M^{-1} b$ and $\widetilde{A} = -A_M^{-1} A_{M^c}$, and in this case we will write $x_M = x_M(y)$.

3.2.2 $K_{k,d}$ boundary classification for the quadratic case and connection with the dual cones

We know that (under proper assumptions, see Theorem 3.1.3) $d(t) \to x^*$ as $t \to \infty$. Naturally, we would like to understand the changes occurring to $(P(d))$ as $d(t) \to x^*$ in order to gain insight into the dynamics of $d(t), x(d(t)), \forall t \geq 0$. In particular, since $x(d) \in \partial K_{k,d} \cap \{x : Ax = b\}$, we are interested in understanding how the boundary changes as $d(t)$ changes.

To start the discussion we consider a simple case when the corresponding wrapping polynomial is just a quadratic form. Suppose $m = 1, n > m$. Therefore, we have $k = 2$ in $(P(d))$ and, for a fixed d, $\widehat{p}(d, x) = (n-2)! E_n(d) E_2(x/d)$. The boundary of $K_{\widehat{p},d}$ is given by

$$E_1(x/d) \geq 0$$
$$E_2\left(\tfrac{x}{d}\right) = 0 \Leftrightarrow \left(\tfrac{x}{d}\right)^T \left[\mathbf{1}\mathbf{1}^T - I\right] \left(\tfrac{x}{d}\right) = 0$$

Introducing $y := [x_2, \ldots, x_{n-1}, x_n] \in \mathbb{R}^{n-1}$ and $e := [d_2, \ldots, d_{n-1}, d_n] \in \mathbb{R}^{n-1}$, and rewriting affine feasible points x as $x_M = \widetilde{b} + \widetilde{A} y$, the boundary of $K_{\widehat{p},d}$ intersected with $\{x : Ax = b\}$ can be written as

$$\frac{\widetilde{b} + \widetilde{A} y}{\widetilde{b} + \widetilde{A} e} + \mathbf{1}^T \begin{pmatrix} y \\ e \end{pmatrix} \geq 0$$

and

$$\begin{aligned}
0 &= \begin{pmatrix} \widetilde{b}+\widetilde{A}y \\ y \end{pmatrix}^T [\mathbf{1}/d][\mathbf{1}\mathbf{1}^T - I][\mathbf{1}/d] \begin{pmatrix} \widetilde{b}+\widetilde{A}y \\ y \end{pmatrix} \\
&= \left(\begin{pmatrix} \widetilde{b} \\ 0 \end{pmatrix} + \begin{pmatrix} \widetilde{A} \\ I \end{pmatrix} y \right)^T [\mathbf{1}/d][\mathbf{1}\mathbf{1}^T - I][\mathbf{1}/d] \left(\begin{pmatrix} \widetilde{b} \\ 0 \end{pmatrix} + \begin{pmatrix} \widetilde{A} \\ I \end{pmatrix} y \right) \\
&= \begin{pmatrix} \widetilde{b} \\ 0 \end{pmatrix}^T \left[\frac{\mathbf{1}}{d}\right] [\mathbf{1}\mathbf{1}^T - I] \left[\frac{\mathbf{1}}{d}\right] \begin{pmatrix} \widetilde{b} \\ 0 \end{pmatrix} + 2 \begin{pmatrix} \widetilde{b} \\ 0 \end{pmatrix}^T \left[\frac{\mathbf{1}}{d}\right] [\mathbf{1}\mathbf{1}^T - I] \left[\frac{\mathbf{1}}{d}\right] \begin{pmatrix} \widetilde{A} \\ I \end{pmatrix} y \\
&+ y^T \begin{pmatrix} \widetilde{A} \\ I \end{pmatrix}^T \left[\frac{\mathbf{1}}{d}\right] [\mathbf{1}\mathbf{1}^T - I] \left[\frac{\mathbf{1}}{d}\right] \begin{pmatrix} \widetilde{A} \\ I \end{pmatrix} y = \gamma + z^T y + \frac{1}{2} y^T Q y
\end{aligned}$$

with

$$\gamma = \begin{pmatrix} \widetilde{b} \\ 0 \end{pmatrix}^T [\mathbf{1}/d][\mathbf{1}\mathbf{1}^T - I][\mathbf{1}/d] \begin{pmatrix} \widetilde{b} \\ 0 \end{pmatrix}$$

$$z = 2 \begin{pmatrix} \widetilde{b} \\ 0 \end{pmatrix}^T [\mathbf{1}/d][\mathbf{1}\mathbf{1}^T - I][\mathbf{1}/e] \begin{pmatrix} \widetilde{A} \\ I_{(n-1)\times(n-1)} \end{pmatrix}$$

$$Q = 2 \begin{pmatrix} \widetilde{A} \\ I_{(n-1)\times(n-1)} \end{pmatrix}^T [\mathbf{1}/d][\mathbf{1}\mathbf{1}^T - I][\mathbf{1}/d] \begin{pmatrix} \widetilde{A} \\ I_{(n-1)\times(n-1)} \end{pmatrix}$$

Depending on the eigenvalues of Q we will get an ellipse, (a branch of) parabola or (a branch of) hyperbola.

Example 3.2.4 ((P) in $\mathbb{R}^3$ with single simplex constraint)**.** Consider a particular LP

$$\min(0,1,1)x$$

$$(1,1,1)x = 3$$

$$x \geq 0$$

The optimal solution is $x^* = [3;0;0]$, and $\widehat{p}(x)$ is $E_2(x/d)$. Switching to the basis of null($[1,1,1]$) corresponding to $(y_1, y_2) := (x_2, x_3)$ (similarly for d), we can write

$x_1 = 3 - y_1 - y_2$, $d_1 = 3 - d_2 - d_3 = 3 - e_1 - e_2$, and thus

$$\begin{aligned}
Q &= 2\begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}^T [\mathbf{1}/d][\mathbf{1}\mathbf{1}^T - I][\mathbf{1}/d]\begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \\
&= \frac{2}{d_1 d_2 d_3}\begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}^T \begin{bmatrix} 0 & d_3 & d_2 \\ d_3 & 0 & d_1 \\ d_2 & d_1 & 0 \end{bmatrix}\begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \\
&= \frac{2}{d_1 d_2 d_3}\begin{pmatrix} -2d_3 & d_1 - (d_2 + d_3) \\ d_1 - (d_2 + d_3) & -2d_2 \end{pmatrix} \\
&= \frac{2}{(3 - e_1 - e_2)e_1 e_2}\begin{pmatrix} -2e_2 & 3 - 2(e_1 + e_2) \\ 3 - 2(e_1 + e_2) & -2e_1 \end{pmatrix}
\end{aligned}$$

For $(e_1, e_2) \in \mathbb{R}^2_{++}$ we are interested in the sign pattern for the eigenvalues of Q (i.e., how many are of the same sign and how many are zeros), because this gives us the shape of the boundary of $K_{2,d} \cap \{x : Ax = b\}$. Note that

$$\begin{aligned}
\det(Q) &= 4e_1e_2 - (3 - 2(e_1 + e_2))^2 = 4e_1e_2 - (9 - 12(e_1 + e_2) + 4(e_1 + e_2)^2) \\
&= 4e_1e_2 - (9 - 12e_1 - 12e_2 + 4e_1^2 + 8e_1e_2 + 4e_2^2) \\
&= -9 + 12(e_1 + e_2) - 4(e_1^2 + e_2^2) - 4e_1e_2 = \lambda_1\lambda_2 < 0
\end{aligned}$$

for small e_1, e_2, where λ_1 and λ_2 are the eigenvalues of Q. That is, for (e_1, e_2) sufficiently close to the origin (recall, this is the optimal LP solution y^*), we necessarily get a branch of hyperbola for the boundary of $K_{\widehat{p},d} \bigcap \{x : Ax = b\}$ in the coordinate system corresponding to y. See Figure 3.3 (three different hyperbolicity directions depicted: $e^1, e^2, e^3 \in \mathbb{R}^2$, with corresponding boundaries in $(P(d))$ of $K_{k,d} \bigcap \{x : Ax = b\}$: K_1, K_2, K_3 and optimal solutions y^1, y^2, y^3).

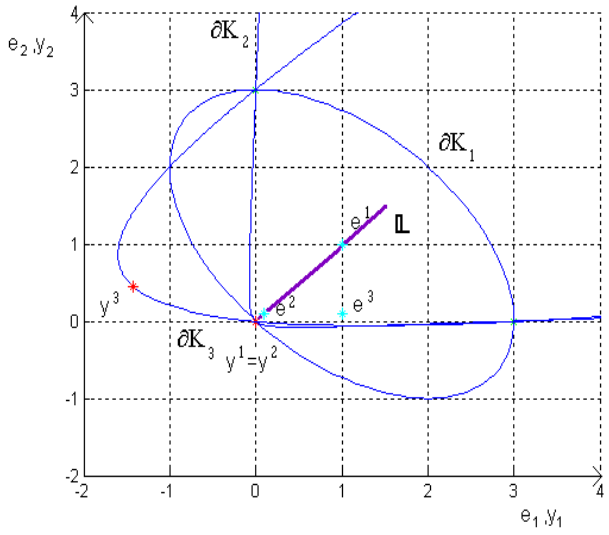

Figure 3.3: Shrink-Wrapping algorithm for LP, relating e^i and K_i

A more general statement can be made telling us when the feasible region of $(P(d))$ can become unbounded:

Proposition 3.2.5. *Consider the following conic programming problem (with a single affine simplex constraint)*

$$(P_s(d)) \quad \{\min_x c^T x : \mathbf{1}^T x = n, x \in K_{k,d}\}$$

for some $d \in \mathbb{R}^n_{++} : \mathbf{1}^T d = n$, $2 \leq k \leq n$. *Then* $(P_s(d))$ *has a bounded feasible region iff* $d \in intK^*_{k,\mathbf{1}}$.

Proof. First note that $(P_s(d))$ is always feasible (take $x_0 = \mathbf{1}$). Therefore our goal is to demonstrate how to find a direction of unboundedness for its feasible domain (find $u \neq 0 : \mathbf{1}^T u = 0, u \in K_{k,d}$) or show that no such direction exists.

Observe that $K_{k,d} = \{[d]x : x \in K_{k,\mathbf{1}}\}$ since $E^{(n-k)}(d, x) = (n-k)!E_n(d)E_k(x/d)$. Since $\mathbb{R}^n_+ \subseteq K_{k,d}$, $K^*_{k,d} \subseteq (\mathbb{R}^n_+)^* = \mathbb{R}^n_+$, the affine constraint $\mathbf{1}^T x = n$ will "capture" all of the dual cone $K^*_{k,d}$ up to scaling (i.e., for any point $y \in K^*_{k,d}, y \neq 0$, there

exists $y_s \in K^*_{k,d} \bigcap \{x : \mathbf{1}^T x = n\}$ s.t. $y = ty_s$ for some $t \geq 0$). Thus, we are not really restricting the choice for d when requiring it to be affine feasible for $(P_s(d))$.

Now suppose $d \notin \text{int}K^*_{k,\mathbf{1}}$. There are two alternatives:

(i) there is $x \in K_{k,\mathbf{1}}$ such that $x^T d < 0$. We can write this as $\mathbf{1}^T[d]x < 0$. We can scale $x_d := [d]x \in K_{k,d}$ to any $tx_d, t > 0$ to still obtain $\mathbf{1}^T(tx_d) < 0$. Also note that surely there is another vector z in $K_{k,d}$ such that $\mathbf{1}^T z > 0$ (e.g., $z = d$, so that $\mathbf{1}^T d > 0$ since $d > 0$). Now we can take a convex combination of these vectors $u := \alpha x_d + (1-\alpha)z, 0 < \alpha < 1$ so that $\mathbf{1}^T u = 0$. Note that $u \in K_{k,d}$ (by convexity), and $u \neq 0$ since $K_{k,d}$ is regular. So u can be scaled to $tu, \forall t \geq 0$, thus giving us a direction of unboundedness for $(P_s(d))$,

(ii) there is $x \in K_{k,\mathbf{1}}, x \neq 0$, such that $x^T d = 0$. Just take $u = [d]x \in K_{k,d}$ to be our direction of unboundedness for $(P_s(d))$.

Conversely, if $d \in \text{int}K^*_{k,\mathbf{1}}$ there is no such vector u and we are done. □

Corollary 3.2.6 (A sufficient condition for $(P(d))$ to have a bounded domain)**.** *Let $(P(d))$ be as before with some $d \in \mathbb{R}^n_{++} : Ad = b$. Suppose $\mathbf{1} \in L^\perp$ and $(P(d))$ is feasible. If $d \in intK^*_{k,\mathbf{1}}$, then $(P(d))$ has a bounded feasible region.*

Proof. Trivially, follows from $(P_s(d))$ being a "relaxed" version of $(P(d))$. □

Remark 3.2.7. The assumption above that $\mathbf{1} \in L^\perp$ is basically same as saying that the original LP (P), if feasible at all, has a bounded feasible region. Note that if $\exists h \in \mathbb{R}^n_{++}$ s.t. $h \in L^\perp$ and the LP is feasible (say, has a feasible point x_0), it means we can add the affine constraint $h^T(x - x_0) = 0$ to the LP without changing its feasible region and by further scaling the coordinates from x to $[1/h]x$ we can assume $h = \mathbf{1}$. Clearly such a constraint will produce a bounded feasible region for

(P) (a subset of a simplex). Conversely, if no such h exists the LP feasible region can be shown to be unbounded, if feasible (e.g., by strong duality for LP).

The proposition and the corollary above could have been established using strong and weak duality for LP, which perhaps would have given a slightly shorter proof. We prefer to present an elementary self-contained argument not relying on duality theory.

Remark 3.2.8. As in Example 3.2.4, in general as $d(t)$ approaches x^*, the resulting relaxation $(P(d))$ will loose the boundedness property for its feasible region except in the special case that the LP solution x^* (or faces, if not unique) are (almost always) in the dual cone $K^*_{k,\mathbf{1}}$.

Corollary 3.2.6 gives us only sufficient condition which is not necessary: consider a similar setting to Example 3.2.4 where $\{x : \mathbf{1}^T x = n\}$ is replaced with

$$x = \begin{pmatrix} 3/2+\gamma \\ 3/2-\gamma \\ 0 \end{pmatrix} \lambda + \begin{pmatrix} 0 \\ 3/2-\gamma \\ 3/2+\gamma \end{pmatrix} (1-\lambda), \forall \lambda \in \mathbb{R}$$

$0 < (-\gamma) < 3/2$ – fixed (set $\gamma = -1/2$, for example).

3.2.3 The central line

In this section we introduce an important geometrical object that proves to be crucial in the analysis of the new algorithmic framework.

We fixed $k = m + 1$. Recall that if $x \in \text{int}K_{m,d}$, we can state the optimality conditions for x to solve $(P(d))$ in terms of the (concave) rational functional

$$q_d(x) := \frac{(n-m)\widehat{p}(d,x)}{(\widehat{p}(d,x))'} = \frac{E_{m+1}(x/d)}{E_m(x/d)}$$

(where $(\widehat{p}(d,\cdot)'$ corresponds to the derivative polynomial of $\widehat{p}(d,\cdot)$ in the direction d; for convenience we slightly augment our previous notation for q_d by a constant

multiplier $(n-m)$). We will simply write $q(x)$ for $q_{\mathbf{1}}(x)$ noting that $q_d(x) = q(x/d)$. Namely, the necessary and sufficient conditions for $x \in \text{int}K_{m,d}$ to be an optimal solution $x(d)$ for $(P(d))$ can be written as

$$q_d(x) = 0$$
$$\nabla_x q_d(x) =_L \tau c$$
$$Ax = b$$

for some $\tau > 0$.

We will show that in fact d can be chosen such that x^* is also a solution to $(P(d))$, that is $x(d) = x^*$, and moreover the following is true.

Theorem 3.2.9 (The "central line"). *There is a line segment (the central line)* $\mathbb{L} \subset \{x \in \mathbb{R}^n : Ax = b, x \in \mathbb{R}^n_{++}\}$ *with* $x^* \in \text{cl}\mathbb{L}$ *s.t. if* $d \in \mathbb{L}$ *then* $x(d) = x^*$. *Moreover, if (w.l.o.g.) we assume* $c_Y := \nabla_y(c^T[x_M(y); y]) = \mathbf{1}$, *then for any* $d \in \mathbb{L}$ *we have* $[d_{m+1}; \ldots; d_n] = t\mathbf{1}$ *for some* $t > 0$.

Proof. Recall our notation $M = \{1, \ldots, m\}$, $M^c = \{1, \ldots, n\} \setminus M$. We assumed that the first m components of x^* are non-zeros. We consider the coordinates for $\{x : Ax = b\}$ that corresponds to the last $(n - m)$ components of x, that is, $y = x_{M^c}$: any affine feasible point x can be written as

$$x = \begin{pmatrix} \widetilde{b} \\ 0 \end{pmatrix} + \begin{pmatrix} \widetilde{A} \\ I \end{pmatrix} y$$

where $x^* = [\widetilde{b}; 0]$.

Consider the gradient $\nabla_x q_d(x)$ at x^*. Obviously, the first m components of $\nabla_x q_d(x^*)$ are zeros, since

$$\nabla_x q_d(x) = (n-m)\nabla_x \left(\frac{\widehat{p}(d,x)}{\widehat{p}(d,x)'} \right) = (n-m) \left(\frac{\nabla_x \widehat{p}(d,x)}{\widehat{p}(d,x)'} - \nabla_x \widehat{p}(d,x)' \frac{\widehat{p}(d,x)}{(\widehat{p}(d,x)')^2} \right)$$

and $\widehat{p}(d, x^*) = 0$, with first m components of $\nabla_x \widehat{p}(d, x^*)$ being zeros as well. Applying the chain rule to compute $\nabla_y q_d([x_M(y); y]) = [\widetilde{A}^T I] \nabla_x q_d(x)|_{x=[x_M(y);y]}$ at $y = 0$ we conclude that

$$\nabla_y q_d([x_M(y); y])|_{y=0} = (\nabla_x q_d(x^*))_{M^c}$$

Our first step is to show that $d : d > 0, Ad = b$, can be chosen such that $x(d) = x^*$. We can write

$$E_{m+1}(x) = E_m(x_M)E_1(y) + E_{m-1}(x_M)E_2(y) + E_{m-2}(x_M)E_3(y) + \cdots$$

$$E_m(x) = E_m(x_M) + E_{m-1}(x_M)E_1(y) + E_{m-2}(x_M)E_2(y) + \cdots$$

Note that $x^* \in \text{int}K_{m,d}$, $\forall d > 0$. Partitioning d into $[d_M(e); e]$ as well, the gradient (w.r.t. y) of the quotient $q_d(\cdot)$ at $y = 0$ (that is at x^*) can be written as

$$\begin{aligned}
\nabla_y q\left(\left[\frac{x_M(y)}{d_M(e)}; \frac{y}{e}\right]\right) &= \frac{1}{E_m\left(\frac{x}{d}\right)}\left(\left[\frac{\mathbf{1}}{e}\right]\mathbf{1}E_m\left(\frac{x_M(y)}{d_M(e)}\right) + \nabla_y E_m\left(\frac{x_M(y)}{d_M(e)}\right)E_1\left(\frac{y}{e}\right)\right. \\
&+ E_{m-1}\left(\frac{x_M(y)}{d_M(e)}\right)\nabla_y E_2\left(\frac{y}{e}\right) \\
&+ \left.\nabla_y E_{m-1}\left(\frac{x_M(y)}{d_M(e)}\right)E_2\left(\frac{y}{e}\right) + \cdots\right) \\
&- \frac{E_{m+1}\left(\frac{x}{d}\right)}{E_m\left(\frac{x}{d}\right)^2}\left(\nabla_y E_m\left(\frac{x_M(y)}{d_M(e)}\right) + E_{m-1}\left(\frac{x_M(y)}{d_M(e)}\right)\nabla_y E_1\left(\frac{y}{e}\right)\right. \\
&+ \left.\nabla_y E_{m-1}\left(\frac{x_M(y)}{d_M(e)}\right)E_1\left(\frac{y}{e}\right) + \cdots\right) \\
&= \left(\frac{\mathbf{1}}{e}\right)
\end{aligned}$$

since $q_d(x^*) = 0$ and $E_m(x/d) = E_m(x_M(y)/d_M(e))$. Given the above, the LP optimal solution x^* will solve $(P(d))$ if

$$\nabla_x q_d(x^*) =_L \tau c$$

for some $\tau > 0$, which is the same as

$$\begin{pmatrix} 0 \\ \nabla_y q_d([x_M(y); y])|_{y=0} \end{pmatrix} =_L \tau c \qquad (3.2.3.1)$$

Observe that $L = \text{null}(A)$ satisfies

$$L = \{x : x = \begin{bmatrix} \widetilde{A} \\ I \end{bmatrix} y, \text{ for some } y\}$$

and consequently

$$L^{\perp} = \text{range}(A) = \{x : x^T \begin{bmatrix} \widetilde{A} \\ I \end{bmatrix} = 0\}$$

Therefore 3.2.3.1 can be rewritten as

$$\left(\begin{pmatrix} 0 \\ z \end{pmatrix} - c \right) \begin{bmatrix} \widetilde{A} \\ I \end{bmatrix} = 0$$

with $z = 1/\tau \left[\frac{\mathbf{1}}{e}\right] \mathbf{1}$ (recall $d = [d_M(e); e]$). Since x^* is the unique optimal solution (the strict minimizer) for (P), it must be that

$$c^T \left(\begin{bmatrix} \widetilde{A} \\ I \end{bmatrix} y \right) > 0, \forall y \geq 0$$

(recall the parametrization of the affine feasible region and the LP feasible region in terms of y). Hence

$$([\widetilde{A}^T I]c) \in \text{int}((\mathbb{R}_+^{n-m})^*) \equiv \mathbb{R}_{++}^{n-m}$$

and as long as we pick $z = ([\widetilde{A}^T I]c)$ the condition 3.2.3.1 will be satisfied. So, finally, choose $e > 0$ satisfying

$$e = \tau \left(\frac{\mathbf{1}}{[\widetilde{A}^T I]c} \right)$$

such that $[d_M(y); y] \in \mathbb{R}_{++}^n$ (just pick e small enough, i.e., pick small enough τ, say $\bar{\tau}$), for $x([d_M(e); e]) = x^*$ to hold.

Note that the optimality conditions for x^* to solve $(P([d_M(e); e]))$ will also be met for any

$$e = \tau \left(\frac{\mathbf{1}}{[\widetilde{A}^T I]c} \right), 0 < \tau < \bar{\tau}$$

thus giving us the central line $\mathbb{L}$.

Regarding generality of the assumption $c_Y = ([\widetilde{A}^T I]c) = \mathbf{1}$, observe that we can re-scale (P) using appropriate $d \in \mathbb{L}$ (e.g., $x \mapsto (x/d)$). If $c_Y = \mathbf{1}$ it is now obvious that $\forall d \in \mathbb{L}, d_{M^c} = t\mathbf{1}$ (for some $t > 0$). □

For simplicity from this point on we assume that for a pair (P), $(P(d))$, the central line corresponds to

$$\mathbb{L} = \{d : d = t\mathbf{1}, 0 < t < \bar{t}\}$$

and consequently

$$c_Y = \nabla y(c^T[x_M(y); y]) = \mathbf{1}$$

3.2.4 The Jacobian of $x(d)$ on the central line

We have just demonstrated that there is an invariant set $\mathbb{L}$ such that if $d \in \mathbb{L}$, then the corresponding $x(d) = x^*$. In order for us to understand the dynamics for $d(t), x(d(t)), t \geq 0$ in the proximity of the optimal LP solution, we first investigate how $x(d)$ changes if we perturb d from the central line $\mathbb{L}$ ever so slightly.

With the help of the rational functional q_d we can understand the behavior of $x(d)$ (up to higher order terms) for small deviations of d from $\mathbb{L}$. Namely, we can compute the Jacobian of $y(e)$, $Dy(e)$, at a point $e \in \mathbb{L}_{M^c}$ (that is, corresponding to the hyperbolicity direction $d = [d_M(e); e] \in \mathbb{L}$). First we need to evaluate the Hessian of q_d.

As before, we write

$$E_{m+1}(x) = E_m(x_M)E_1(y) + E_{m-1}(x_M)E_2(y) + E_{m-2}(x_M)E_3(y) + \cdots$$
$$E_m(x) = E_m(x_M) + E_{m-1}(x_M)E_1(y) + E_{m-2}(x_M)E_2(y) + \cdots$$

Consider the Hessian of $q_d(\cdot)$ with respect to y for affine feasible $x = [x_M(y); y]$ at $y = 0$ (that is at x^*) for (affine feasible) $d = [d_M(e); e]$. We have

$$\begin{aligned}
\nabla^2_{yy} q\left(\left[\tfrac{x_M(y)}{d_M(e)}; \tfrac{y}{e}\right]\right) &= \frac{\nabla^2_{yy} E_{m+1}\left(\left[\frac{x_M(y)}{d_M(e)}; \frac{y}{e}\right]\right)}{E_m\left(\left[\frac{x_M(y)}{d_M(e)}; \frac{y}{e}\right]\right)} \\
&- \frac{1}{E_m\left(\left[\frac{x_M(y)}{d_M(e)}; \frac{y}{e}\right]\right)^2} \nabla_y E_{m+1}\left(\left[\tfrac{x_M(y)}{d_M(e)}; \tfrac{y}{e}\right]\right) \nabla_y E_m\left(\left[\tfrac{x_M(y)}{d_M(e)}; \tfrac{y}{e}\right]\right)^T \\
&- \frac{1}{E_m\left(\left[\frac{x_M(y)}{d_M(e)}; \frac{y}{e}\right]\right)^2} \nabla_y E_m\left(\left[\tfrac{x_M(y)}{d_M(e)}; \tfrac{y}{e}\right]\right) \nabla_y E_{m+1}\left(\left[\tfrac{x_M(y)}{d_M(e)}; \tfrac{y}{e}\right]\right)^T \\
&+ \frac{2E_{m+1}\left(\left[\frac{x_M(y)}{d_M(e)}; \frac{y}{e}\right]\right)}{E_m\left(\left[\frac{x_M(y)}{d_M(e)}; \frac{y}{e}\right]\right)^3} \nabla_y E_m\left(\left[\tfrac{x_M(y)}{d_M(e)}; \tfrac{y}{e}\right]\right) \nabla_y E_m\left(\left[\tfrac{x_M(y)}{d_M(e)}; \tfrac{y}{e}\right]\right)^T \\
&- \frac{E_{m+1}\left(\left[\frac{x_M(y)}{d_M(e)}; \frac{y}{e}\right]\right)}{E_m\left(\left[\frac{x_M(y)}{d_M(e)}; \frac{y}{e}\right]\right)^2} \nabla^2_{yy} E_m\left(\left[\tfrac{x_M(y)}{d_M(e)}; \tfrac{y}{e}\right]\right) \\
&= \frac{1}{E_m\left(\frac{x}{d}\right)} \nabla_y \left(\left[\tfrac{\mathbf{1}}{e}\right] \mathbf{1} E_m\left(\tfrac{x_M(y)}{d_M(e)}\right)\right. \\
&+ \nabla_y E_m\left(\tfrac{x_M(y)}{d_M(e)}\right) E_1\left(\tfrac{y}{e}\right) + E_{m-1}\left(\tfrac{x_M(y)}{d_M(e)}\right) \nabla_y E_2\left(\tfrac{y}{e}\right) \\
&+ \left.\nabla_y E_{m-1}\left(\tfrac{x_M(y)}{d_M(e)}\right) E_2\left(\tfrac{y}{e}\right) + \cdots\right) \\
&- \frac{1}{E_m\left(\frac{x_M(y)}{d_M(e)}\right)^2} E_m\left(\tfrac{x_M(y)}{d_M(e)}\right)\left(\tfrac{\mathbf{1}}{e}\right)\left(\nabla_y E_m\left(\tfrac{x_M(y)}{d_M(e)}\right)\right. \\
&+ \left. E_{m-1}\left(\tfrac{x_M(y)}{d_M(e)}\right)\left(\tfrac{\mathbf{1}}{e}\right)\right)^T \\
&- \frac{1}{E_m\left(\frac{x_M(y)}{d_M(e)}\right)^2} \left(\nabla_y E_m\left(\tfrac{x_M(y)}{d_M(e)}\right)\right. \\
&+ \left. E_{m-1}\left(\tfrac{x_M(y)}{d_M(e)}\right)\left(\tfrac{\mathbf{1}}{e}\right)\right) E_m\left(\tfrac{x_M(y)}{d_M(e)}\right)\left(\tfrac{\mathbf{1}}{e}\right)^T
\end{aligned}$$

and thus

$$\begin{aligned}
\nabla^2_{yy} q_d|_{y=0} &= \frac{1}{E_m\left(\frac{x_M(y)}{d_M(e)}\right)} \left(\nabla_y E_m\left(\tfrac{x_M(y)}{d_M(e)}\right)\left(\tfrac{\mathbf{1}}{e}\right)^T + \left(\tfrac{\mathbf{1}}{e}\right) \nabla_y E_m\left(\tfrac{x_M(y)}{d_M(e)}\right)^T\right. \\
&+ \left. E_{m-1}\left(\tfrac{x_M(y)}{d_M(e)}\right)\left[\tfrac{\mathbf{1}}{e}\right]\left(\mathbf{1}\mathbf{1}^T - I\right)\left[\tfrac{\mathbf{1}}{e}\right]\right) \\
&- \frac{1}{E_m\left(\frac{x_M(y)}{d_M(e)}\right)} \left(\left(\tfrac{\mathbf{1}}{e}\right) \nabla_y E_m\left(\tfrac{x_M(y)}{d_M(e)}\right)^T + \left(\tfrac{\mathbf{1}}{e}\right) E_{m-1}\left(\tfrac{x_M(y)}{d_M(e)}\right)\left(\tfrac{\mathbf{1}}{e}\right)^T\right. \\
&+ \left. \nabla_y E_m\left(\tfrac{x_M(y)}{d_M(e)}\right)\left(\tfrac{\mathbf{1}}{e}\right)^T + \left(\tfrac{\mathbf{1}}{e}\right) E_{m-1}\left(\tfrac{x_M(y)}{d_M(e)}\right)\left(\tfrac{\mathbf{1}}{e}\right)^T\right) \\
&= \frac{E_{m-1}}{E_m}\left(\frac{x^*_M}{d_M(e)}\right)\left[\tfrac{\mathbf{1}}{e}\right]\left(-I - \mathbf{1}\mathbf{1}^T\right)\left[\tfrac{\mathbf{1}}{e}\right]
\end{aligned}$$

Knowing the Hessian of q_d with respect to y at $y = 0$ (at LP optimum x^*), we can proceed to analyze the derivative (Jacobian) of $y(e)$ for $e \in \mathbb{L}_{M^c}$. We will differentiate the implicit function defining $y(e)$.

Recall $q(x) = q_{\mathbf{1}}(x)$ and we assumed $c_Y = \mathbf{1}$, so that $\exists \bar{t} > 0$ such that $\forall t \in (0, \bar{t})$, $e = t\mathbf{1}$ gives us the hyperbolicity direction $d = [d_M(e); e]$ corresponding to $x(d) = x^*$. The optimality conditions for $(P(d))$, assuming $[x_M(y); y] \in \text{int}K_{m,d}$,

$$q_d([x_M(y); y]) = 0$$
$$\nabla_y q_d([x_M(y); y]) = \tau c_Y$$

can be written as

$$\begin{aligned} q(x)|_{\left(\frac{x}{d}\right)} &= 0 \\ [\widetilde{A}^T I] \left[\tfrac{1}{d}\right] \nabla_x q(x)|_{\left(\frac{x}{d}\right)} &= \tau \mathbf{1} \end{aligned} \tag{3.2.4.1}$$

Remark 3.2.10 (On the value of τ for $d \in \mathbb{L}$). Observe that from evaluation of the gradient of q_d with respect to y at $y = 0$ (i.e., at x^*) with $d \in \mathbb{L}$ ($e = t\mathbf{1}, t > 0$) and the optimality conditions above

$$\nabla_y q_d([x_M(y); y]) = \left[\frac{1}{e}\right] \mathbf{1} = \tau \mathbf{1}$$

Pre-multiplying by $e^T = t\mathbf{1}^T$,

$$(n - m) = t\mathbf{1}^T \left(\frac{1}{t}\right) \mathbf{1} = \tau t \mathbf{1}^T \mathbf{1} = \tau t (n - m)$$

It necessarily follows that $\tau = 1/t$.

Noting that the solution to the set of equations 3.2.4.1, giving us optimality conditions for $y = y(e)$, is itself a function of e, as is $\tau = \tau(e)$, differentiating with respect to e, and for compactness denoting $x = [x_M(y(e)); y(e)])$, we have

$$\begin{aligned} &\left(\nabla_x q(x)|_{\left(\frac{x}{d}\right)}\right)^T \left[\tfrac{1}{d}\right] \begin{bmatrix} \widetilde{A} \\ I \end{bmatrix} \dot{y} + \left(\nabla_x q(x)|_{\left(\frac{x}{d}\right)}\right)^T \left[\tfrac{-x}{d^2}\right] \begin{bmatrix} \widetilde{A} \\ I \end{bmatrix} = 0 \\ &\qquad -[\widehat{A}^T I] \left[\tfrac{1}{d}\right] \left[\nabla_x q(x)|_{\left(\frac{x}{d}\right)}\right] \left[\tfrac{1}{d}\right] \begin{bmatrix} \widetilde{A} \\ I \end{bmatrix} \\ &+[\widehat{A}^T I] \left[\tfrac{1}{d}\right] \left(\nabla^2_{xx} q(x)|_{\left(\frac{x}{d}\right)} \left[\tfrac{1}{d}\right] \begin{bmatrix} \widetilde{A} \\ I \end{bmatrix} \dot{y} - \nabla^2_{xx} q(x)|_{\left(\frac{x}{d}\right)} \left[\tfrac{-x}{d^2}\right] \begin{bmatrix} \widetilde{A} \\ I \end{bmatrix} \right) = \mathbf{1}\dot{\tau} \end{aligned} \tag{3.2.4.2}$$

where $\dot{y}$ is a Jacobian matrix of $y(e)$ ($\dot{y}_{i,j} = \frac{\partial y_i}{\partial e_j}$), and $\dot{\tau}$ is a first derivative row-vector of τ with respect to e ($\dot{\tau}_j = \frac{\partial \tau}{\partial e_j}$).

Observe at $x = [x_M(y); y]$ (by the chain rule)

$$\nabla^2_{yy} q_d([x_M(y); y]) = [\widehat{A}^T I] \begin{bmatrix} \frac{1}{d} \end{bmatrix} \nabla^2_{xx} q(x)|_{(\frac{x}{d})} \begin{bmatrix} \frac{1}{d} \end{bmatrix} \begin{bmatrix} \widetilde{A} \\ I \end{bmatrix}$$

so that at $y = 0$ (i.e., at x^*) we have

$$[\widehat{A}^T I] \begin{bmatrix} \frac{1}{d} \end{bmatrix} \nabla^2_{xx} q(x)|_{(\frac{x^*}{d})} \begin{bmatrix} \frac{1}{d} \end{bmatrix} \begin{bmatrix} \widetilde{A} \\ I \end{bmatrix} = \frac{E_{m-1}}{E_m} \left(\frac{x^*_M}{d_M(e)} \right) \begin{bmatrix} \frac{1}{e} \end{bmatrix} (-I - \mathbf{1}\mathbf{1}^T) \begin{bmatrix} \frac{1}{e} \end{bmatrix}$$

Also

$$\begin{aligned} \frac{1}{n-m} \nabla^2_{xx} q_d(x) &= \left[\frac{\nabla^2_{xx} \widehat{p}}{\widehat{p}'} - \frac{\nabla_x \widehat{p} (\nabla_x \widehat{p}')^T}{(\widehat{p}')^2} - \frac{\nabla_x \widehat{p}' (\nabla_x \widehat{p})^T}{(\widehat{p}')^2} \right. \\ &- \left. \nabla^2_{xx} \widehat{p}' \frac{\widehat{p}}{(\widehat{p}')^2} + \frac{2\widehat{p}}{(\widehat{p}')^3} \nabla_x \widehat{p}' (\nabla_x \widehat{p}')^T \right] (d, x) \end{aligned}$$

(we slightly abuse notation here by carrying the parameter d and the argument x of the polynomials $\widehat{p}$, $\widehat{p}'$, outside of the brackets surrounding the whole expression). At $x = x^*$, and more generally, at any $x \geq 0$ having precisely m first coordinates non-zero, $\widehat{p}(d, x) = 0$, $\widehat{p}'(d, x) > 0$, and the first m components of $\nabla_x \widehat{p}(d, x)$ are zeros together with $(\nabla^2_{xx} \widehat{p}(d, x))_{i,j} = 0, \forall i, j \leq m$. Hence we conclude that

$$(\nabla^2_{xx} q_d(x))_{i,j} = 0, \forall i, j \leq m$$

(in particular, at $y = 0$) and

$$\frac{1}{n-m} \nabla^2_{xx} q_d(x) = \left[\frac{\nabla^2_{xx} \widehat{p}}{\widehat{p}'} - \frac{\nabla_x \widehat{p} (\nabla_x \widehat{p}')^T}{(\widehat{p}')^2} - \frac{\nabla_x \widehat{p}' (\nabla_x \widehat{p})^T}{(\widehat{p}')^2} \right] (d, x)$$

Also, if $x \geq 0$ is a point having precisely m first nonzero coordinates we have

$$\nabla_x q(x) = \begin{pmatrix} 0 \\ \nabla_y q(x) \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{1} \end{pmatrix}$$

W.l.o.g. (for simplicity) for a moment we may assume that the point e, at which we are evaluating the derivative $\dot{y}(e)$, corresponds to the hyperbolicity direction $d = \mathbf{1} \in \mathbb{L}$ (by strict feasibility of d). Then

$$q_d(x) = q(x) = \frac{E_{m+1}(x)}{E_m(x)} = (n-m)\frac{\widehat{p}(x)}{(\widehat{p}(x))'}$$

and at x^*

$$\nabla_x q(x^*) = \begin{pmatrix} 0 \\ \mathbf{1} \end{pmatrix}$$

where

$$\frac{1}{(n-m)}\nabla q(x^*) = \frac{\nabla_x \widehat{p}(x^*)}{(\widehat{p}(x^*))'}$$

In this case we can write

$$\begin{aligned} \frac{1}{(n-m)}\nabla^2_{xx} q(x^*) &= \frac{\nabla^2_{xx}\widehat{p}(x^*)}{\widehat{p}'(x^*)} - \left(\frac{1}{(n-m)}\right) \\ &\left(\begin{pmatrix} 0 \\ \mathbf{1} \end{pmatrix}\left(\frac{\nabla_x \widehat{p}'(x^*)}{\widehat{p}'(x^*)}\right)^T + \left(\frac{\nabla_x \widehat{p}'(x^*)}{\widehat{p}'(x^*)}\right)\begin{pmatrix} 0 \\ \mathbf{1} \end{pmatrix}^T\right) \end{aligned}$$

and since $\widehat{p}(x) = \widehat{p}(\mathbf{1}, x) = (n-m-1)!E_{m+1}(x)$, $\widehat{p}'(x) = \widehat{p}'(\mathbf{1}, x) = (n-m-1)!(n-m)E_m(x)$, we have

$$\left(\nabla^2_{xx} q(x^*)\right)_{i,j} = 0, \forall i, j \notin M^c = \{m+1, \ldots, n\}$$

(by comparing the terms in the Hessian resulting from differentiation of these elementary symmetric functions evaluated at x^*). This line of reasoning can be easily extended to the case of arbitrary $d > 0, d \in \mathbb{L}$. As a straightforward consequence of this we get

$$\nabla^2_{xx} q_d(x^*) = \begin{bmatrix} 0 & 0 \\ 0 & \nabla^2_{yy} q_d(x^*) \end{bmatrix}$$

With this in mind, we can rewrite 3.2.4.2 for $d \in \mathbb{L}$, that is $e = d_{M^c} = t\mathbf{1}, t > 0$, as follows

$$\left(\tfrac{1}{e}\right)\dot{y} = 0$$
$$-\left[\tfrac{1}{e}\right] I \left[\tfrac{1}{e}\right] + \left[\tfrac{1}{e}\right]\left(\tfrac{E_{m-1}}{E_m}\left(\tfrac{x_M^*}{d_M(e)}\right)(-I - \mathbf{1}\mathbf{1}^T)\left[\tfrac{1}{e}\right]\dot{y}\right) = \mathbf{1}\dot{\tau}$$

that is

$$\tfrac{1}{t}\mathbf{1}^T\dot{y} = 0$$
$$\tfrac{1}{t^2}\left(-I + \tfrac{E_{m-1}}{E_m}\left(\tfrac{x_M^*}{d_M(e)}\right)(-I - \mathbf{1}\mathbf{1}^T)\dot{y}\right) = \mathbf{1}\dot{\tau}$$

so the first equation says that the columns of $\dot{y}$ must be orthogonal to $\mathbf{1}$. From the second equation we get

$$\dot{y} = \frac{E_m}{E_{m-1}}\left(\frac{x_M^*}{d_M(e)}\right)\left(-I + \frac{\mathbf{1}\mathbf{1}^T}{1+(n-m)}\right)(t^2\mathbf{1}\dot{\tau} + I)$$

We want to evaluate $\dot{\tau}$. The orthogonality condition becomes

$$\mathbf{1}^T\left(-I + \frac{\mathbf{1}\mathbf{1}^T}{1+(n-m)}\right)(t^2\mathbf{1}\dot{\tau} + I) = 0$$

or considering τ_i separately (for any fixed i)

$$\mathbf{1}^T\left(-I + \frac{\mathbf{1}\mathbf{1}^T}{1+(n-m)}\right)(t^2\mathbf{1}\dot{\tau}_i + I_i) = 0$$

(with I_i being an i^{th} column of identity matrix). We get

$$\begin{aligned} 0 &= t^2\dot{\tau}_i\left((n-m) - \frac{(n-m)^2}{1+(n-m)}\right) + \left(1 - \frac{(n-m)}{1+(n-m)}\right) \\ &= t^2\dot{\tau}_i\frac{(n-m)}{1+(n-m)} + \frac{1}{1+(n-m)} \end{aligned}$$

so

$$\dot{\tau}_i = -1/(t^2(n-m))$$

Finally

$$\begin{aligned} \dot{y} &= \frac{E_m\left(\frac{x_M^*}{d_M(e)}\right)}{E_{m-1}\left(\frac{x_M(y)}{d_M(e)}\right)}\left(-I + \frac{\mathbf{1}\mathbf{1}^T}{1+(n-m)}\right)\left(\frac{-1}{(n-m)}\mathbf{1}\mathbf{1}^T + I\right) \\ &= \frac{E_m}{E_{m-1}}\left(\frac{x_M^*}{d_M(e)}\right)\left(-I + \frac{\mathbf{1}\mathbf{1}^T}{n-m}\right) \end{aligned}$$

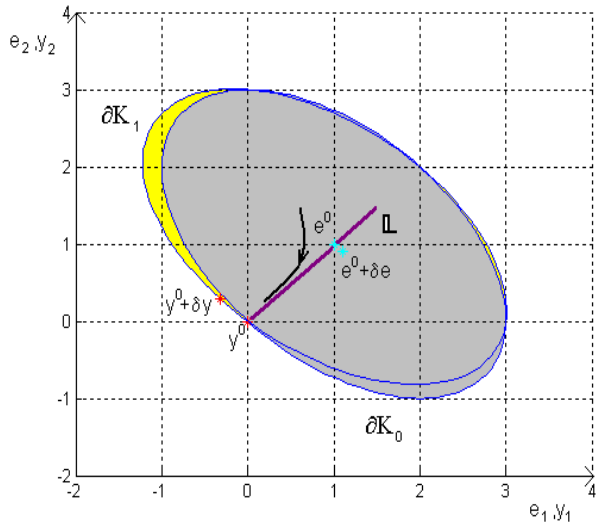

Figure 3.4: Shrink-Wrapping algorithm for LP, derivative of $y(e)$

(at $d = [d_M(e); e] \in \mathbb{L}$, i.e., $e = t\mathbf{1}, t > 0$). Note that under our assumptions for LP this quantity has a finite limit as $t \downarrow 0$ for $e = t\mathbf{1}$ (think of a "derivative" of $y(e)$ as $e \to 0$). Obviously, the (positive) multiplier will tend to $\frac{1}{m}$ (both elementary symmetric polynomials will be evaluated at a vector of all ones).

Interpretation: one way to write the first differential of $y(e)$ at $e = t\mathbf{1} \in \mathbb{L}_{M^c}, t > 0$ (the reference point being on the central line) would be:

$$Dy(e) : \delta e \mapsto \frac{E_m}{E_{m-1}} \left(\frac{x_M^*}{d_M(e)} \right) \left(-\mathrm{proj}_{\mathrm{null}(\mathbf{1}^T)}(\delta e) \right)$$

Noting that $\frac{E_m}{E_{m-1}} \left(\frac{x_M^*}{d_M(e)} \right)$ must stay positive (as a ratio functional inside the smaller of the two cones) we can say that small deviations of y (or d) from the central line (measured in terms of orthogonal distance to $\mathbb{L}_{M^c}$) are "counteracted" (up to some positive multiplier) by the corresponding shifts of $y(e)$ from $y = 0$ (if $y(e)$ is a "reasonable" function, e.g., $y(e) \in C^2$). See Figure 3.4 (same setting as in Example 3.2.4). Here we depict two distinct hyperbolicity directions $e^0 \in$

$\mathbb{L}_{M^c}, e^1 = e^0 + \delta e$ with corresponding boundaries of the feasible regions for $(P(d))$: $\partial K_0, \partial K_1$, the optimal solution $y^0 = y(e^0)$ and the approximation of $y^1 = y(e^1)$ with its first differential $y^0 + \delta y$ (the difference will be indistinguishable in the figure, so only the approximation is shown).

3.3 Understanding the dynamics close to the central line

The central line $\mathbb{L}$ was defined as a set of hyperbolicity directions such that $\forall d \in \mathbb{L}$, $x(d) = x^*$. So as long as $d(t), t \geq 0$ (given by 3.1.0.1) "locates" this invariant set $\mathbb{L}$ fairly quickly, we would expect $x(d(t))$ to converge to x^* quickly as well.

Our goal is to construct a (discrete) algorithm that would have nice convergence properties for the LP (P). It turns out that fast (super-linear, quadratic) convergence of the iterates $\{x_i\}_{i\geq 0}$ is intrinsically connected to the geometrical nature of the convergence of the iterates $\{d_i\}_{i\geq 0}$ to the central line $\mathbb{L}$, and therefore understanding the nature of convergence in the continuous setting for $d(t), t \geq 0$, is crucial (namely we want to exhibit "asymptotic" convergence to $\mathbb{L}$, see Figure 3.4, the black trajectory with an arrow depicts the targeted shape for $\{e(t), t \geq 0\}$).

Recall we assumed (w.l.o.g.) that $c_Y = \mathbf{1}$. In what follows (see subsection 3.3.1) we will establish the

Theorem 3.3.1 (Limiting behavior of $e(t)$ as $t \uparrow \infty$)**.** *There is a wedge*

$$W_{e_0,\epsilon} = \{e \in \mathbb{R}^{n-m} : e = t(e_0 + \delta e), 0 \leq t \leq 1, \delta e \in \text{null}(\mathbf{1}^T), \|\delta e\| \leq \epsilon\}$$

of the central line $\mathbb{L}_{M^c} (\ni e_0 = \tau\mathbf{1}, \tau > 0)$ (with $\epsilon > 0$) such that for any starting point in $W_{e_0,\epsilon}$, the solution to the ODE 3.2.1.1

$$\dot{e} = y(e) - e$$

will stay in $W_{e_0,\epsilon}$*, converge asymptotically to* $\mathbb{L}$

$$\frac{\|e_\perp(t)\|}{e_\parallel(t)} \to 0$$

and exponentially fast (if $\|e_\perp(0)\| \neq 0$*), meaning*

$$\frac{\|e_\perp(t)\|}{e_\parallel(t)} \sim C_1 e^{-\frac{1}{m}t}$$

as $t \uparrow \infty$ *(for some* $C_1 > 0$*), where* $e_\parallel(t) = \text{proj}_{\text{range}(\mathbf{1}^T)} e(t)$*,* $e_\perp(t) = \text{proj}_{\text{null}(\mathbf{1}^T)} e(t)$ *(by* $f(t) \sim g(t)$ *for* $t \uparrow \infty$ *we mean* $\frac{f(t)}{g(t)} \to 1$*).*

The proof will give us a better understanding of the relationship between $e(t)$ and $y(e(t))$ (in a certain neighborhood of x^*); this will provide us with means to quickly construct and analyze a (locally) super-linear convergent algorithm for LP.

Remark 3.3.2. If at $y = 0$ (i.e., at x^*) we could simply approximate $y(e)$ with its linearization (note that the only possible candidate for this is the limit of the Jacobian of $y(e)$ as $e = t\mathbf{1}, t \downarrow 0$) then the desired asymptotic convergence ($\frac{\|e_\perp(t)\|}{e_\parallel(t)} \to 0$ as $t \uparrow \infty$) would follow from standard ODE results (e.g., by looking at solutions to the linearized ODE $\dot{e} = \left(-\frac{1}{m}\left(I - \frac{\mathbf{1}\mathbf{1}^T}{n-m}\right) - I\right) e$ in a neighborhood of a stationary point $e = 0$). However in a later subsection (see 3.3.2) we will

- demonstrate why this (standard) argument does not apply for our case
- propose a way to overcome this problem and demonstrate that under proper assumptions we will indeed have the desired conclusion regarding $\{e(t), t \geq 0\}$

3.3.1 Euclidian coordinates approach

The ODE for $e(t), t \geq 0$, is given by 3.2.1.1 where $y(e)$ (in general) is the solution to a non-linear system of algebraic (e.g., polynomial) equations. Our goal is to

understand how $e(t)$ acts in a certain neighborhood of the central line $\mathbb{L}_{M^c}$ close to the optimal LP solution x^* (i.e., $e = 0$).

The main difficulty to be overcome is gaining a better understanding of $y(e)$ as a function of e. In order to do so, we rely on approximating $y(e)$ (for $e \in W_{e_0,\epsilon}$, referred to as proper neighborhood of $e = 0$) in two steps.

We can first approximate $y(e)$ – a solution to a system of nonlinear equations (see Remark 3.3.3 below)

$$g(e;y) := \begin{pmatrix} \text{proj}_{\text{null}\left(c_Y^T\right)} \nabla_y q_{[d_M(e);e]}([x_M(y);y]) \\ q_{[d_M(e);e]}([x_M(y);y]) \end{pmatrix} = 0$$

with first iterate of Newton's method, $N_g(e, y_0)$ (we treat e as a fixed parameter), corresponding to the fixed starting point $y_0 = 0$, which in turn can be approximated by the first differential of $y(e)$, $Dy(e)|_{e_0}\delta e$, with $e = e_0 + \delta e$, $e_0 \in \mathbb{L}_{M^c}$, $\delta e \in \mathbb{L}_{M^c}^{\perp}$,

$$y(e) \approx N_g(e, y_0) \approx Dy(e)|_{e_0}\delta e$$

See Figure 3.5 (same setting as in the Example 3.2.4).

In analyzing the quality of these approximations, namely, how well the Newton's iterate approximates the true solution to $g(e;y) = 0$, we, to a large extent, rely on the Newton's method analysis presented in [19] (see Appendix B).

Remark 3.3.3 (On usage of g(e;y) in characterization of $y(e)$). Recall that $g(e;y) = 0$ would give us necessary and sufficient conditions for $y = y(e)$ iff coupled with $[x_M(y);y] \in \text{int}K_{m,[d_M(e);e]}$ (i.e., $E_{m-i}(x/d) > 0, 0 \leq i \leq (k-1)$), and $\tau c_Y = \nabla_y q_{[d_M(e);e]}([x_M(y);y])$ with some $\tau > 0$ (for otherwise we could potentially arrive at the maximizer instead of the minimizer).

From the analysis of Newton's method it will follow that for e sufficiently close to e_0, $y \mapsto g(e;y)$ is real analytic (in a sense of convergent power series) in the

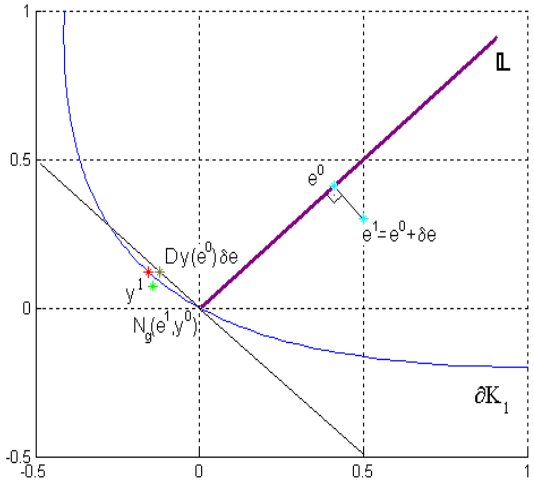

Figure 3.5: Shrink-Wrapping algorithm for LP, approximating $y(e)$

ball centered at $y_0 = 0$ of radius $\|2N_g(e, y_0)\|$ with the corresponding root of $g(e; y) = 0$ being in this ball. In particular $q_{[d_M(e);e]}([x_M(y); y])$ has to be finite and thus $\widehat{p}'(d, [x_M(y); y]) \neq 0$ in this ball, so the rest of the constraints, $[x_M(y); y] \in \text{int}K_{k-1,[d_M(e);e]}$, can be discarded in the characterization of $y(e)$ (see the note in the proof of Theorem 3.3.10).

Analogously, one can argue that if e is sufficiently close to e_0, then the corresponding solution to $g(e; y) = 0$ will satisfy $q_{[d_M(e);e]}([x_M(y); y]) = \tau c_Y$ for some $\tau > 0$ as well (see the same note as mentioned above).

Two key ingredients that go into proof of Theorem 3.3.1 are:

- understanding the behavior of a certain system of differential equations of first order,
- finding "good" error bounds on the approximation of $y(e)$ by $Dy(e_0)\delta e$.

First exercise in ODE

We start with analyzing ODE in $\mathbb{R}^2$ of the form

$$\begin{cases} \dot{x}_1 &= -x_1 + (\eta + O_1(|x_1| + |x_2|))O_{2,1}\left(\frac{x_2^2}{x_1}\right) \\ & \quad + O_{3,1}\left(\frac{x_2^2}{x_1}\right) =: f_1(x_1, x_2, t) \\ \dot{x}_2 &= -x_2 + (\eta + O_1(|x_1| + |x_2|))(-x_2) \\ & \quad + (\eta + O_1(|x_1| + |x_2|))O_{2,2}\left(\frac{x_2^2}{x_1}\right) + O_{3,2}\left(\frac{x_2^2}{x_1}\right) =: f_2(x_1, x_2, t) \end{cases} \tag{3.3.1.1}$$

with the initial conditions (at $t_0 = 0$) given by

$$x_1(0) = \bar{x}_1, x_2(0) = \bar{x}_2 \tag{3.3.1.2}$$

$\bar{x}_1 = 1, |\bar{x}_2|$ being possibly $\ll 1$ and $0 < \eta$. (We denote the solution to 3.3.1.1, 3.3.1.2 as $(x_1(t), x_2(t))$ sometimes referring to a particular value at time t as $(x_1, x_2)_t$.)

Suppose f_1, f_2 are continuous on $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 \geq 0\}$, so that the solution to this initial value problem exists.

Remark 3.3.4. We are not going to be concerned with the uniqueness of the solution to this initial value problem (recall one sufficient condition for uniqueness is local Lipschitz continuity of f_1, f_2 with respect to x on our domain; if no Lipschitz continuity is assumed, one might have non-unique solutions in general: e.g., consider $\dot{x} = 2|x|^{1/2}, x(0) = 0$, two solutions $x(t) \equiv 0, x(t) = t^2\text{sign}(t)$). The existence statement (implied by continuity of f) will suffice for our purposes.

We want to understand the behavior of the system as $t \uparrow \infty$ under some additional assumptions on $O_i, O_{i,j} : \mathbb{R} \to \mathbb{R}$. In particular, $|O_1(y)| \leq K_1|y|, |O_{i,j}(y)| \leq K_i|y|$ (Lipschitz continuous at 0) with some constants K_i. While we impose no assumptions on the magnitude of K_2, K_3, we will suppose that K_1 is relatively small with respect to η, for example $K_1 < \frac{\eta}{2}$.

Remark 3.3.5 (Why do we choose this particular ODE?). Note that $y(e)$ (or $x(d)$) is continuous, since the boundary of $K_{m+1,d}$ has strict curvature outside of $\mathbb{R}^n_+$ and changes continuously as e (or $d = [d_M(e); e]$) changes, so for our ODE 3.2.1.1, its right-hand side, $(y(e) - e)$, is continuous as well.

For a moment assume $e, y \in \mathbb{R}^2, e = e_\parallel + e_\perp$, with the coordinate system chosen such that $e_\parallel$ and $e_\perp$ are aligned with the coordinate axis, $e_1 = \|e_\parallel\|$, $|e_2| = \|e_\perp\|$. Assume for a moment that the first Newton iterate (with e fixed) satisfies

$$\begin{aligned} N_g(e,0) &= \left(\frac{1}{m} + O_1(|e_1| + |e_2|)\right) \begin{pmatrix} 0 \\ \\ -e_2 \end{pmatrix} + \mathbf{1}O_2\left(\frac{e_2^2}{e_1}\right) \\ &\approx [Dy(e_\parallel)]e_\perp = \frac{E_m}{E_{m-1}}\left(\frac{x_M^*}{d_M(e_\parallel)}\right) \begin{pmatrix} 0 \\ \\ -e_2 \end{pmatrix} \\ &= \left(\frac{1}{m} + O_1(|e_1| + |e_2|)\right) \begin{pmatrix} 0 \\ \\ -e_2 \end{pmatrix} \approx \frac{1}{m} \begin{pmatrix} 0 \\ \\ -e_2 \end{pmatrix} \end{aligned}$$

and furthermore

$$y(e) = N_g(e,0) + \mathbf{1}O_3\left(\frac{e_2^2}{e_1}\right)$$

(if, say, $e_1 \equiv 1$, we would expect the Newton step to give us a "quadratic error" $\approx O(e_2^2)$, recalling $N_g(e,0) \approx -\frac{1}{m}e_\perp$). In such a case 3.3.1.1, 3.3.1.2 is what we want to understand (up to scaling) in order to understand the behavior of $e(t)$ as $t \uparrow \infty$.

Intuitively, one can look at system 3.3.1.1 as

$$\begin{cases} \dot{x}_1 \approx -x_1 \\ \\ \dot{x}_2 \approx (-1-\eta)x_2 \end{cases}$$

so one would expect (if the approximation is "well-behaved") the solution to be of the form $x_1(t) \approx \bar{x}_1 e^{-t}, x_2(t) \approx \bar{x}_2 e^{(-1-\eta)t}$. The goal of the following discussion is to make this statement precise.

Lemma 3.3.6. *Suppose $\eta > 0, K_1 < \eta/2$. Then the constant $\epsilon > 0$ can be chosen such that for any starting point $(\bar{x}_1, \bar{x}_2)$ in a wedge*

$$W_\epsilon := \left\{(x_1, x_2) \in \mathbb{R}^2 : 0 \leq x_1 \leq 1, 0 \leq |x_2| \leq \epsilon x_1\right\}$$

the solution to 3.3.1.1, 3.3.1.2 will approach the origin, staying in W_ϵ, $\forall t \geq 0$. Moreover, if $x_2(0) \neq 0$, then as $t \uparrow \infty$

$$x_1(t) \sim C_1 e^{-t}$$
$$x_2(t) \sim C_2 e^{-(1+\eta)t}$$

for some constants C_1, C_2.

Proof. Consider the solution to 3.3.1.1, 3.3.1.2 corresponding to the initial point $(\bar{x}_1, \bar{x}_2)$ in a "flat disk"

$$\omega_\epsilon := \left\{(x_1, x_2) \in \mathbb{R}^2 : x_1 = 1, |x_2| \leq \epsilon\right\}$$

As we will show, the solution to 3.3.1.1, 3.3.1.2 does not change signs componentwise, so (w.l.o.g.) assume $\bar{x}_2 > 0$. (The case of $\bar{x}_2 = 0$ is very simple: note that if $x_2(0) = 0$ initially, then as a direct consequence of Lipschitz continuity we get $x_2(t) \equiv 0$, $\forall t \geq 0$, and therefore $x_1(t) = e^{-t}$ follows.)

Since the solution to 3.3.1.1, 3.3.1.2 is continuous, given some $\alpha, \beta > 1$, we can pick $\Delta t > 0$ small enough such that $x_1(t) \in [1/\alpha, \beta]$, $x_2(t) \geq 0$ on $[0, \Delta t]$. We will impose some conditions on β a bit later.

From Lipschitz continuity of O_1 at 0 it follows that

$$|O_1(|x_1| + |x_2|)| \leq K_1(|x_1| + |x_2|) \leq K_1\beta + K_1 x_2$$

on $[0, \Delta t]$. Also

$$\begin{aligned} f_2(x_1(0), x_2(0), 0) &= -\bar{x}_2 + (\eta + O_1(\bar{x}_1 + \bar{x}_2))(-\bar{x}_2) \\ &+ (\eta + O_1(\bar{x}_1 + \bar{x}_2))O_{2,2}\left(\frac{\bar{x}_2^2}{\bar{x}_1}\right) + O_{3,2}\left(\frac{\bar{x}_2^2}{\bar{x}_1}\right) \\ &\leq -\bar{x}_2 - \eta\bar{x}_2 + \bar{x}_2(1K_1 + K_1\bar{x}_2) \\ &+ (\eta + 1K_1 + K_1\bar{x}_2)K_2\bar{x}_2^2 + K_3\bar{x}_2^2 \\ &= \bar{x}_2(-1 - \eta + K_1 + K_1\bar{x}_2) \\ &+ \bar{x}_2^2((\eta + K_1 + K_1\bar{x}_2)K_2 + K_3) < 0 \end{aligned}$$

provided $\bar{x}_2$ is small enough. Assume we start with such $\bar{x}_2$.

Consequently $x_2(t)$ will decrease in some (possibly even smaller than $[0, \Delta t]$) neighborhood of $t_0 = 0$, say $[0, \tau]$. Therefore, for $t \in [0, \Delta t] \bigcap [0, \tau]$, we have

$$|f_2(x_1, x_2, t) - (-1 - \eta)x_2| \leq \gamma x_2 \tag{3.3.1.3}$$

where

$$\gamma := K_1\beta + K_1(x_2)_0 + (\eta + \beta K_1 + K_1(x_2)_0)K_2\alpha(x_2)_0 + K_3\alpha(x_2)_0$$

Assume $\beta \leq 2$ and $K_1 < \eta/2$. Then by choosing $\bar{x}_2$ (i.e., ϵ) sufficiently small, we may also assume

$$-1 - \eta + \gamma < 0 \tag{3.3.1.4}$$

We now demonstrate 3.3.1.4 ensures that $x_2(t)$ is decreasing in the whole of $[0, \Delta t]$, so the bound 3.3.1.3 holds uniformly on $[0, \Delta t]$, and thus there is no need to reduce the time interval to $[0, \tau] \bigcap [0, \Delta t]$.

Proposition 3.3.7 (A slightly refined version of Comparison Lemma, e.g., [20])**.** *Suppose we have two ODE's of the form*

$$\dot{x} = h_1(x, t) \qquad \dot{y} = h_2(y, t)$$

with h_1, h_2 being continuous with respect to t on some interval $[0, \Delta t]$, h_2 locally Lipschitz continuous with respect to the first argument (with t fixed in $[0, \Delta t]$), and h_1 continuous with respect to the first argument. Given the same initial condition $x_0 = y_0$, suppose that the solutions to both initial value problems exist and are decreasing on the time interval $[0, \Delta t]$; denote these $x(t), y(t)$. Moreover, suppose that $h_1(y(t), t) < h_2(y(t), t)$ for all $t \in [0, \Delta t]$. Then

$$x(t) \leq y(t), \forall t \in [0, \Delta t]$$

Proof. Note that by assumptions imposed on h_1, h_2, we are guaranteed to have solutions $x(t), y(t)$, and moreover, we have a uniqueness of the solution $y(t)$ for $t_0 \in [0, \Delta t]$.

Both $x(t)$ and $y(t)$ continuous, so (by the mean-value theorem) $x(t) < y(t)$ in some small neighborhood of $0 : t > 0$. In other words, initially the solution curve $x(t)$ goes underneath $y(t)$. Now suppose at some point $t^* \in [0, \Delta t]$ these curves cross again, that is $x(t^*) = y(t^*)$, then necessarily we must have $x'(t^*) \geq y'(t^*)$, which implies $h_1(t^*, y(t^*)) \geq h_2(t^*, y(t^*))$, contradiction. □

As a consequence of the proposition above and 3.3.1.3 we have

$$\bar{x}_2 e^{(-1-\eta-\gamma)t} \leq x_2(t) \leq \bar{x}_2 e^{(-1-\eta+\gamma)t} \tag{3.3.1.5}$$

on $[0, \Delta t]$. Note that, strictly speaking, we should have a strict inequality in 3.3.1.3 in order to apply the proposition, but it is easily seen that one can employ a limiting argument, first throwing in an extra term, say of the form εx_2, into 3.3.1.3 and then letting $\varepsilon \downarrow 0$ to obtain 3.3.1.5.

Now let us consider $x_1(t)$. From

$$f_1(x_1, x_2, t) = -x_1 + (\eta + O_1(|x_1| + |x_2|))O_{2,1}\left(\frac{x_2^2}{x_1}\right) + O_{3,1}\left(\frac{x_2^2}{x_1}\right)$$

and the bound 3.3.1.5 on $x_2(t)$ on $[0, \Delta t]$ we get

$$|f_1(x_1, x_2, t) + x_1| \leq ((\eta + K_1\beta + K_1\bar{x}_2\alpha K_2 + \alpha K_3)\bar{x}_2^2 e^{2(-1-\eta+\gamma)t}$$

Moreover, letting

$$\xi := ((\eta + K_1\beta + K_1\bar{x}_2)\alpha K_2 + \alpha K_3)\bar{x}_2^2$$

we claim that if

$$-1 - \xi > 2(-1 - \eta + \gamma) \tag{3.3.1.6}$$

then for $t \in [0, \Delta t]$,

$$\bar{x}_1 e^{(-1-\xi)t} \leq x_1(t) \leq \bar{x}_1 e^{(-1+\xi)t} \tag{3.3.1.7}$$

Indeed, these bounds also follow from the proposition above and the following observation. Suppose we have the equation with $0 < \mu < 1$

$$\dot{x} = -x - \mu e^{-2t}, x(0) = 1$$

and we compare its solution $x(t)$ with the one of

$$\dot{y} = (-1 - \mu)y, y(0) = 1$$

that is, $y(t) = e^{(-1-\mu)t}$. Obviously

$$-y(t) - \mu e^{-2t} = -e^{(-1-\mu)t} - \mu e^{-2t} \geq (-1 - \mu)y(t) = (-1 - \mu)e^{(-1-\mu)t}$$

since

$$-\mu e^{-2t} \geq -\mu e^{(-1-\mu)t}$$

for $t \geq 0$. So $x(t) \geq y(t)$ by the proposition (strictly speaking, we also need to use the limiting argument here). The other bound is established in a similar fashion.

So far, for 3.3.1.7 to hold we require 3.3.1.6: $-1-\xi > 2(-1-\eta+\gamma)$, that is

$$-1-((\eta+K_1\beta+K_1\bar{x}_2)\alpha K_2+\alpha K_3)\bar{x}_2^2$$
$$>2(-1-\eta+((K_1\beta+K_1\bar{x}_2)+(\eta+\beta K_1+K_1\bar{x}_2)K_2\alpha\bar{x}_2+K_3\alpha\bar{x}_2))$$

and observe that again, with $\beta \leq 2$, $K_1 < \eta/2$ this condition is easily met for sufficiently small $\bar{x}_2 > 0$ (i.e., $\epsilon > 0$ small enough).

Naturally, we would prefer $x_1(t)$ also to be a decreasing function on $[0, \Delta t]$. For this it is enough to require $\xi < 1$, that is

$$((\eta+K_1\beta+K_1\bar{x}_2)\alpha K_2+\alpha K_3)\bar{x}_2^2 < 1 \tag{3.3.1.8}$$

again guaranteed for $\bar{x}_2$ small enough. In turn, this shows that the bound on β is somewhat artificial.

To summarize what we have so far: under the conditions 3.3.1.4, 3.3.1.6, 3.3.1.8 (all of which can be simultaneously satisfied by picking $\bar{x}_2$ below a certain threshold ϵ) we have the following bounds on $x_1(t), x_2(t)$ for $t \in [0, \Delta t]$: 3.3.1.5, 3.3.1.7. Moreover, both components of the solution are guaranteed to decrease on this interval and such an interval necessarily exists due to continuity of the solution of 3.3.1.1, 3.3.1.2.

We have established exponential bounds on $x_1(t), x_2(t)$ for $t \in [0, \Delta t]$. It remains to show that, in fact, similar bounds will hold uniformly for all $t \geq 0$.

In order to do so consider a re-scaling of 3.3.1.1, 3.3.1.2. Given $(x_1)_{\Delta t}, (x_2)_{\Delta t}$, we put 3.3.1.1, 3.3.1.2 to its "original" setting corresponding to a starting point at $t_0 = \Delta t$.

Introduce

$$\widetilde{x}_1 := \frac{x_1}{(x_1)_{\Delta t}}, \widetilde{x}_2 := \frac{x_2}{(x_2)_{\Delta t}}\bar{x}_2$$

and let us observe what will happen to $O_i, O_{i,j}$ under this transformation.

Note that

$$\frac{(\widetilde{x}_2)^2}{\widetilde{x}_1} = \frac{x_2^2}{x_1}(x_1)_{\Delta t}\left(\frac{\bar{x}_2}{(x_2)_{\Delta t}}\right)^2$$

Consider an equivalent system to 3.3.1.1

$$\left\{\begin{array}{rcl} \frac{1}{(x_1)_{\Delta t}}\dot{x}_1 & = & \frac{1}{(x_1)_{\Delta t}}\left(-x_1+(\eta+O_1(|x_1|+|x_2|))O_{2,1}\left(\frac{x_2^2}{x_1}\right)\right. \\ & & \left.+O_{3,1}\left(\frac{x_2^2}{x_1}\right)\right) \\ \frac{\bar{x}_2}{(x_2)_{\Delta t}}\dot{x}_2 & = & \frac{\bar{x}_2}{(x_2)_{\Delta t}}\left(-x_2+(\eta+O_1(|x_1|+|x_2|))(-x_2)\right. \\ & & \left.+(\eta+O_1(|x_1|+|x_2|))O_{2,2}\left(\frac{x_2^2}{x_1}\right)+O_{3,2}\left(\frac{x_2^2}{x_1}\right)\right) \end{array}\right. \tag{3.3.1.9}$$

This can be rewritten as

$$\left\{\begin{array}{rcl} \dot{\widetilde{x}}_1 & = & -\widetilde{x}_1+(\eta+O_1(|x_1|+|x_2|))O_{2,1}\left(\frac{(\widetilde{x_2})^2}{\widetilde{x_1}}\frac{1}{(x_1)_{\Delta t}}\left(\frac{(x_2)_{\Delta t}}{\bar{x}_2}\right)^2\right)\frac{1}{(x_1)_{\Delta t}} \\ & & +O_{3,1}\left(\frac{(\widetilde{x_2})^2}{\widetilde{x_1}}\frac{1}{(x_1)_{\Delta t}}\left(\frac{(x_2)_{\Delta t}}{\bar{x}_2}\right)^2\right)\frac{1}{(x_1)_{\Delta t}} \\ \dot{\widetilde{x}}_2 & = & -\widetilde{x}_2+(\eta+O_1(|x_1|+|x_2|))(-\widetilde{x}_2) \\ & & +(\eta+O_1(|x_1|+|x_2|))O_{2,2}\left(\frac{(\widetilde{x_2})^2}{\widetilde{x_1}}\frac{1}{(x_1)_{\Delta t}}\left(\frac{(x_2)_{\Delta t}}{\bar{x}_2}\right)^2\right)\frac{\bar{x}_2}{(x_2)_{\Delta t}} \\ & & +O_{3,2}\left(\frac{(\widetilde{x_2})^2}{\widetilde{x_1}}\frac{1}{(x_1)_{\Delta t}}\left(\frac{(x_2)_{\Delta t}}{\bar{x}_2}\right)^2\right)\frac{\bar{x}_2}{(x_2)_{\Delta t}} \end{array}\right. \tag{3.3.1.10}$$

We assumed that $|O_1(y)| \leq K_1|y|, |O_{i,j}(y)| \leq K_i|y|$, therefore

$$\begin{array}{rcl} \left|O_{2,1}\left(\frac{(\widetilde{x_2})^2}{\widetilde{x_1}}\frac{1}{(x_1)_{\Delta t}}\left(\frac{(x_2)_{\Delta t}}{\bar{x}_2}\right)^2\right)\frac{1}{(x_1)_{\Delta t}}\right| & \leq & K_2\left|\frac{(\widetilde{x_2})^2}{\widetilde{x_1}}\right|\left(\frac{1}{(x_1)_{\Delta t}}\left(\frac{(x_2)_{\Delta t}}{\bar{x}_2}\right)^2\right)\frac{1}{(x_1)_{\Delta t}} \\ & \leq & K_2\left|\frac{(\widetilde{x_2})^2}{\widetilde{x_1}}\right| \end{array}$$

provided

$$\left(\frac{1}{(x_1)_{\Delta t}}\left(\frac{(x_2)_{\Delta t}}{\bar{x}_2}\right)\right) \leq 1$$

Similarly

$$\begin{array}{rcl} \left|O_{3,2}\left(\frac{(\widetilde{x_2})^2}{\widetilde{x_1}}\frac{1}{(x_1)_{\Delta t}}\left(\frac{(x_2)_{\Delta t}}{\bar{x}_2}\right)^2\right)\frac{\bar{x}_2}{(x_2)_{\Delta t}}\right| & \leq & K_3\left|\frac{(\widetilde{x_2})^2}{\widetilde{x_1}}\right|\left(\frac{1}{(x_1)_{\Delta t}}\left(\frac{(x_2)_{\Delta t}}{\bar{x}_2}\right)^2\frac{\bar{x}_2}{(x_2)_{\Delta t}}\right) \\ & \leq & K_3\left|\frac{(\widetilde{x_2})^2}{\widetilde{x_1}}\right| \end{array}$$

again provided

$$\left(\frac{1}{(x_1)_{\Delta t}}\left(\frac{(x_2)_{\Delta t}}{\bar{x}_2}\right)\right) \leq 1$$

Finally, $O_1(|x_1|+|x_2|) = O_1\left(|\widetilde{x}_1|(x_1)_{\Delta t} + |\widetilde{x}_2|\frac{(x_2)_{\Delta t}}{\bar{x}_2}\right)$, and so is Lipschitz continuous at 0 with respect to $|\widetilde{x}_1| + |\widetilde{x}_2|$ with at most the same Lipschitz constant as K_1 (since $x_1(t), x_2(t)$ are decreasing), and, in fact, the new Lipschitz constant will decrease exponentially compared to the "original" K_1 (where the degree of exponent will correspond to the slowest possible decay rate in 3.3.1.7, 3.3.1.5).

So if we can guarantee

$$\left(\frac{1}{(x_1)_{\Delta t}}\left(\frac{(x_2)_{\Delta t}}{\bar{x}_2}\right)\right) \leq 1 \tag{3.3.1.11}$$

we can rewrite 3.3.1.10 as

$$\begin{cases} \dot{\widetilde{x}}_1 &= -\widetilde{x}_1 + (\eta + O_1(|\widetilde{x}_1| + |\widetilde{x}_2|))O_{2,1}\left(\frac{\widetilde{x}_2^2}{\widetilde{x}_1}\right) \\ & \quad + O_{3,1}\left(\frac{\widetilde{x}_2^2}{\widetilde{x}_1}\right) \\ \dot{\widetilde{x}}_2 &= -\widetilde{x}_2 + (\eta + O_1(|\widetilde{x}_1| + |\widetilde{x}_2|))(-\widetilde{x}_2) \\ & \quad + (\eta + O_1(|\widetilde{x}_1| + |\widetilde{x}_2|))O_{2,2}\left(\frac{\widetilde{x}_2^2}{\widetilde{x}_1}\right) + O_{3,2}\left(\frac{\widetilde{x}_2^2}{\widetilde{x}_1}\right) \end{cases} \tag{3.3.1.12}$$

with all of the $O_1, O_{i,j}$ again being Lipschitz continuous at 0 with constants $\widetilde{K_i} \leq K_i$, and moreover $\widetilde{K_1} < K_1$, decaying exponentially fast over $[0, \Delta t]$. The initial conditions 3.3.1.2 are "the same" for system 3.3.1.12; now ($(\widetilde{x}_1)_{\Delta t} = 1$, $|(\widetilde{x}_2)_{\Delta t}| \leq |\bar{x}_2|$).

Note that the re-scaling condition 3.3.1.11 gives us precisely the wedge W_ϵ for all the solutions to 3.3.1.1, 3.3.1.2 with starting point $(\bar{x}_1, \bar{x}_2) \in \omega_\epsilon$ with $|\bar{x}_2|$ being below some threshold value ϵ. It also tells us that any system 3.3.1.1 with initial point in W_ϵ can be re-scaled to produce 3.3.1.12, and thus we can apply to it the argument above.

We can iterate the argument above, letting $t \uparrow \infty$, provided 3.3.1.11 is satisfied. To check that this condition can be met, observe

$$\left(\frac{1}{(x_1)_{\Delta t}}\left(\frac{(x_2)_{\Delta t}}{\bar{x}_2}\right)\right) \leq \frac{\bar{x}_2 e^{(-1-\eta+\gamma)\Delta t}}{\bar{x}_2 \bar{x}_1 e^{(-1-\xi)\Delta t}} = e^{(-\eta+\gamma+\xi)\Delta t}$$

and is ≤ 1 if we require that

$$\begin{aligned}(-\eta+\gamma+\xi) &= -\eta + ((K_1\beta + K_1\bar{x}_2) \\ &+ (\eta + \beta K_1 + K_1\bar{x}_2)K_2\alpha\bar{x}_2 + K_3\alpha\bar{x}_2) \\ &+ ((\eta + K_1\beta + K_1\bar{x}_2)\alpha K_2 + \alpha K_3)\bar{x}_2^2 < 0\end{aligned} \tag{3.3.1.13}$$

which again can be satisfied with a proper (sufficiently small) choice of $|\bar{x}_2|$ (i.e., ϵ).

Finally, let us examine how good are the established bounds in 3.3.1.5, 3.3.1.7 as $t \uparrow \infty$. On $[0, \Delta t]$ we have

$$\bar{x}_2 e^{(-1-\eta-\gamma)t} \leq x_2(t) \leq \bar{x}_2 e^{(-1-\eta+\gamma)t}$$
$$\bar{x}_1 e^{(-1-\xi)t} \leq x_1(t) \leq \bar{x}_1 e^{(-1+\xi)t}$$

Now if we choose $\bar{x}_2$ to have strict inequality

$$(-1-\eta+\gamma) < (-1-\xi)$$

which again, is possible, then the ratio $\frac{x_2(t)}{x_1(t)}$ will also decay exponentially over $[0, \Delta t]$. Thus by iterating the bounds, we will arrive at progressively smaller $(\widetilde{x}_2)_{\Delta t}$ compared to $(\widetilde{x}_1)_{\Delta t} = 1$. Recalling that $\widetilde{K_1}$ also decays exponentially on each interval of length Δt, we eventually obtain our asymptotic result

$$x_1(t) \sim C_1 e^{-t}$$
$$x_2(t) \sim C_2 e^{-(1+\eta)t}$$

□

Computing the Newton step at x^* (i.e., at $y = 0$)

Recall that for $\{x : Ax = b\}$ we adapted the coordinate system corresponding to last $(n - m)$ components of x, denoted $y = x_{M^c}$ (with $M = \{1, \ldots, m\}, M^c = \{1, \ldots, n\} \setminus M$). Similarly, for any affine feasible hyperbolicity direction d we introduced $e = d_{M^c}$ (so that $d = [d_M(e); e]$).

For a given d (determined by e) the first order necessary and sufficient optimality conditions determining the corresponding $x(d)$, or rather $y(e)$, are given by the following system of equations (assuming $[x_M(y); y] \in \text{int}K_{m,[d_M(e);e]}$, see later note in the proof of Theorem 3.3.10)

$$\begin{cases} \nabla_y q_d([x_M(y); y]) = \tau \mathbf{1} \\ q_d([x_M(y); y]) = 0 \end{cases}$$

for some $\tau > 0$, which means that y is a solution of

$$g(e; y) = \begin{pmatrix} \left(I - \frac{\mathbf{1}\mathbf{1}^T}{n-m}\right) \nabla_y q_d([x_M(y); y]) \\ q_d([x_M(y); y]) \end{pmatrix} = 0$$

(recall $c_Y = \mathbf{1}$).

Let us compute the first Newton iterate, $N_g(e, y)$, for some (fixed) e starting at the point $y = 0$ (corresponding to x^*). We need to resolve the linearized (with respect to y) version of $g(e, y) = 0$ to find the Newton step from $y = 0$, $NS_g(e, 0)$ (writing $N_g(e, y) = y + NS_g(e, y)$ using the notation of [19]), that is, $NS_g(e, 0)$ must satisfy

$$\begin{cases} \left(I - \frac{\mathbf{1}\mathbf{1}^T}{n-m}\right) (\nabla_y q_d([x_M(y); y])|_{y=0} + \nabla^2_{yy} q_d([x_M(y); y])|_{y=0} NS_g(e, 0)) = 0 \\ q_d([x_M(y); y])|_{y=0} + \nabla_y q_d([x_M(y); y])|^T_{y=0} NS_g(e, 0) = 0 \end{cases}$$

which can be rewritten as

$$\begin{cases} \nabla_y q_d([x_M(y); y])|_{y=0} + \nabla^2_{yy} q_d([x_M(y); y])|_{y=0} NS_g(e, 0) = \tau \mathbf{1} \\ q_d([x_M(y); y])|_{y=0} + \nabla_y q_d([x_M(y); y])|^T_{y=0} NS_g(e, 0) = 0 \end{cases}$$

for some $\tau > 0$. Thus the Newton step is determined by

$$\begin{cases} \left(\frac{E_{m-1}}{E_m}\left(\frac{x_M^*}{d_M}\right)\right)\left[\frac{1}{e}\right](-I - \mathbf{1}\mathbf{1}^T)\left[\frac{1}{e}\right] NS_g(e,0) = \tau\mathbf{1} - \left[\frac{1}{e}\right]\mathbf{1} \\ \left[\frac{1}{e}\right]\mathbf{1}^T NS_g(e,0) = 0 \end{cases}$$

(with $d_M = d_M(e)$). From the first equation we get

$$NS_g(e,0) = \left(\frac{E_m}{E_{m-1}}\left(\frac{x_M^*}{d_M}\right)\right)[e](I + \mathbf{1}\mathbf{1}^T)^{-1}[e]\left(\left[\frac{1}{e}\right]\mathbf{1} - \tau\mathbf{1}\right)$$

and we can use this to compute τ

$$\begin{aligned} 0 &= \left[\frac{1}{e}\right]\mathbf{1}^T NS_g(e,0) \\ &= \left[\frac{1}{e}\right]\mathbf{1}^T\left(\frac{E_m}{E_{m-1}}\left(\frac{x_M^*}{d_M}\right)\right)[e](I + \mathbf{1}\mathbf{1}^T)^{-1}[e]\left(\left[\frac{1}{e}\right]\mathbf{1} - \tau\mathbf{1}\right) \\ &= \left(\frac{E_m}{E_{m-1}}\left(\frac{x_M^*}{d_M}\right)\right)\mathbf{1}^T(I + \mathbf{1}\mathbf{1}^T)^{-1}[e]\left(\left[\frac{1}{e}\right]\mathbf{1} - \tau\mathbf{1}\right) \\ &= \left(\frac{E_m}{E_{m-1}}\left(\frac{x_M^*}{d_M}\right)\right)\mathbf{1}^T(I + \mathbf{1}\mathbf{1}^T)^{-1}(\mathbf{1} - \tau e) \end{aligned}$$

Hence

$$\tau = \frac{\mathbf{1}^T(I + \mathbf{1}\mathbf{1}^T)^{-1}\mathbf{1}}{\mathbf{1}^T(I + \mathbf{1}\mathbf{1}^T)^{-1}e} = \frac{\mathbf{1}^T\left(I - \frac{\mathbf{1}\mathbf{1}^T}{1+(n-m)}\right)\mathbf{1}}{\mathbf{1}^T\left(I - \frac{\mathbf{1}\mathbf{1}^T}{1+(n-m)}\right)e} = \frac{\mathbf{1}^T\mathbf{1}}{\mathbf{1}^T e}$$

Thus we can write

$$NS_g(e,0) = \left(\frac{E_m}{E_{m-1}}\left(\frac{x_M^*}{d_M(e)}\right)\right)[e]\left(I - \frac{\mathbf{1}\mathbf{1}^T}{1 + (n-m)}\right)(\mathbf{1} - \tau e)$$

with

$$\tau = \frac{\mathbf{1}^T\mathbf{1}}{\mathbf{1}^T e}$$

For the purpose of our analysis it is more convenient to work with an approximate Newton step which we compute as follows.

Let $e = d_{M^c}$ be decomposed into $e_\parallel$ and $e_\perp$, where $e_\parallel = e_{\text{range}(\mathbf{1}^T)} = \frac{e^T\mathbf{1}}{n-m}\mathbf{1}$ (the component corresponding to the central line $\mathbb{L}_{M^c}$), $e = e_\parallel + e_\perp$. Note $|e_\perp| < e_\parallel$ (componentwise) since e must lie in $\mathbb{R}^{(n-m)}_{++}$.

Denoting

$$\alpha = \left(\frac{E_m}{E_{m-1}}\left(\frac{x_M^*}{d_M(e)}\right)\right)$$

we can write

$$\begin{aligned} NS_g(e,0) &= \alpha[e_\| + e_\perp]\left(I - \frac{\mathbf{1}\mathbf{1}^T}{1+(n-m)}\right)\left(\mathbf{1} - \frac{(n-m)}{\mathbf{1}^T e_\|}(e_\| + e_\perp)\right) \\ &= \alpha[e_\| + e_\perp]\left(\mathbf{1} - \frac{(n-m)}{1+(n-m)}\mathbf{1} - \frac{(n-m)}{\mathbf{1}^T e_\|}e_\|\right. \\ &+ \left.\frac{(n-m)}{1+(n-m)}\mathbf{1} - \frac{(n-m)}{\mathbf{1}^T e_\|}e_\perp\right) \\ &= \alpha[e_\| + e_\perp]\left(\mathbf{1} - \frac{(n-m)}{(\mathbf{1}^T e_\|)}(e_\| + e_\perp)\right) \end{aligned}$$

Letting $e_\| = t\mathbf{1}$ we get

$$\begin{aligned} NS_g(e,0) &= \alpha[t\mathbf{1} + e_\perp]\left(\mathbf{1} - \tfrac{(n-m)}{(\mathbf{1}^T\mathbf{1}t)}(t\mathbf{1} + e_\perp)\right) \\ &= \alpha[t\mathbf{1} + e_\perp]\left(\mathbf{1} - \tfrac{(n-m)t}{(n-m)t}\mathbf{1} - \tfrac{(n-m)}{(n-m)}\tfrac{e_\perp}{t}\right) \\ &= \alpha[t\mathbf{1} + e_\perp]\left(-\tfrac{e_\perp}{t}\right) = \alpha\left(-e_\perp - \tfrac{(e_\perp)^2}{t}\right) \end{aligned} \tag{3.3.1.14}$$

Finally we can write

$$\begin{aligned} NS_g(e,0) &= \left(\tfrac{E_m}{E_{m-1}}\left(\tfrac{x_M^*}{d_M(e)}\right)\right)\left(-e_\perp - \tfrac{(e_\perp)^2}{e_\|}\right) \\ &= \left(\tfrac{E_m}{E_{m-1}}\left(\tfrac{x_M^*}{d_M(e)}\right)\right)(-e_\perp + V) \end{aligned} \tag{3.3.1.15}$$

where

$$\|V\|_\infty = O\left((n-m)\frac{\|e_\perp\|^2}{\|e_\|\|}\right)$$

with $|O(y)| \leq |y|$.

Note that for $\alpha = \left(\frac{E_m}{E_{m-1}}\left(\frac{x_M(y)}{d_M(e)}\right)\right)$, at $y = e = 0$ we have $x_M(0) = d_M(0) = x_M^* \neq 0$ (by the non-degeneracy assumption), giving us $\alpha = \frac{1}{m}$. Also, $d_M(e)$ is a linear function of e, so in some small neighborhood of $e = 0$, α as a function of e (with fixed $y = 0$) is Lipschitz continuous. Thus in this neighborhood we can write

$$\left|\alpha(e) - \frac{1}{m}\right| \leq K\|e\| \tag{3.3.1.16}$$

where $\|\cdot\|$ can be an arbitrary norm in $\mathbb{R}^{(n-m)}$ (by equivalence of norms in finite-dimensional vector space), $K > 0$ being a constant depending on the norm. In particular, for $e = e_{\|} + e_{\perp}$ we define $\|e\| = \|e_{\|}\|_2 + \|e_{\perp}\|_2$ where $\|\cdot\|_2$ is the Euclidean norm. We will see later why this choice is useful.

Understanding higher derivatives of $q(x)$

For a moment, consider a (multi-)rational functional

$$q = \frac{p}{f} : \mathbb{R}^n \to \mathbb{R}$$

where $p : \mathbb{R}^n \to \mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$ are multivariate polynomials, and let $m \geq \max(\deg(p), \deg(f))$. In particular, we are interested in understanding the higher-order derivatives of $q(x)$ at a point z such that $f(z) \neq 0$.

Write

$$q(x) = p(x)\frac{1}{f(x)}$$

Differentiating by parts and using the definition of a norm for a multilinear operator $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n \equiv \mathbb{R}^{n^k} \to \mathbb{R}$

$$\|\mathcal{L}\| = \sup_{u_1,u_2,\ldots,u_k,\|u_i\|\leq 1} |\mathcal{L}(u_1, u_2, \ldots, u_k)|$$

(which can be defined using arbitrary norm in $\mathbb{R}^n$; for concreteness we take $\|\cdot\| = \|\cdot\|_2$ – the Euclidean norm) we can write

$$\|q^{(k)}\| \leq \sum_{i=0}^{k} \begin{pmatrix} k \\ i \end{pmatrix} \left\| \left(\frac{1}{f}\right)^{(i)} \right\| \|p^{(k-i)}\|$$

(Note that for some $q = p \cdot g : \mathbb{R}^n \to \mathbb{R}$ we can write

$$D^{(m)}q(x)(V^1, V^2, \ldots, V^m) = \sum_{\substack{i_1=1,\ldots,n \\ i_2=1,\ldots,n \\ \vdots \\ i_m=1,\ldots,n}} \frac{\partial^m}{\partial x_{i_1}\partial x_{i_2}\cdots\partial x_{i_m}} q(x) V^1_{i_1} V^2_{i_2} \cdots V^m_{i_m}$$

where $V^i \in \mathbb{R}^n, i = 1, \ldots, m$. Also,

$$\begin{aligned}\frac{\partial^m}{\partial x_{i_1}\partial x_{i_2}\cdots\partial x_{i_m}} q(x)V_{i_1}^1 V_{i_2}^2\cdots V_{i_m}^m &= \sum_{j=0}^{m} \sum_{\substack{\{h_1<h_2<\cdots<h_j\}\bigcup\{k_{j+1}<k_{j+2}<\cdots<k_m\}\in P(m)\\ \{h_1<h_2<\cdots<h_j\}\bigcap\{k_{j+1}<k_{j+2}<\cdots<k_m\}=\emptyset}} \\ &\quad \frac{\partial^j}{\partial x_{i_{h_1}}\partial x_{i_{h_2}}\cdots\partial x_{i_{h_j}}} p(x)V_{i_{h_1}}^{h_1}V_{i_{h_2}}^{h_2}\cdots V_{i_{h_m}}^{h_m} \\ &\quad \frac{\partial^{(m-j)}}{\partial x_{i_{k_{j+1}}}\partial x_{i_{k_{j+2}}}\cdots\partial x_{i_{k_m}}} g(x)V_{i_{k_{j+1}}}^{k_{j+1}}V_{i_{k_{j+2}}}^{k_{j+2}}\cdots V_{i_{k_m}}^{k_m}\end{aligned}$$

where $P(m)$ is the set of all possible permutations of $\{1, 2, \ldots, m\}$. If $j = 0$, we interpret

$$\frac{\partial^j}{\partial x_{i_{h_1}}\partial x_{i_{h_2}}\cdots\partial x_{i_{h_j}}} p(x)V_{i_{h_1}}^{h_1}V_{i_{h_2}}^{h_2}\cdots V_{i_{h_j}}^{h_j}$$

as simply $p(x)$, and similarly, for $j = m$ we interpret the 0^{th} partial derivative of g as $g(x)$. Substituting and changing the order of summation we write

$$\begin{aligned}D^{(m)}q(x)(V^1, V^2, \ldots, V^m) &= \sum_{j=0}^{m} \sum_{\substack{\{h_1<h_2<\cdots<h_j\}\bigcup\{k_{j+1}<k_{j+2}<\cdots<k_m\}\in P(m)\\ \{h_1<h_2<\cdots<h_j\}\bigcap\{k_{j+1}<k_{j+2}<\cdots<k_m\}=\emptyset}} \\ &\quad \sum_{\substack{i_1=1,\ldots,n\\ i_2=1,\ldots,n\\ \vdots\\ i_m=1,\ldots,n}} \frac{\partial^j}{\partial x_{i_{h_1}}\partial x_{i_{h_2}}\cdots\partial x_{i_{h_j}}} p(x)V_{i_{h_1}}^{h_1}V_{i_{h_2}}^{h_2}\cdots V_{i_{h_m}}^{h_j} \\ &\quad \frac{\partial^{(m-j)}}{\partial x_{i_{k_{j+1}}}\partial x_{i_{k_{j+2}}}\cdots\partial x_{i_{k_m}}} g(x)V_{i_{k_{j+1}}}^{k_{j+1}}V_{i_{k_{j+2}}}^{k_{j+2}}\cdots V_{i_{k_m}}^{k_m}\end{aligned}$$

Regrouping the terms and restricting ourselves to $V^i \in \mathbb{R}^n$ such that $\|V^i\| \leq 1, \forall i$,

we obtain

$$\begin{aligned} D^{(m)}q(x)(V^1,V^2,\ldots,V^m) &\leq \sum_{j=0}^{m} \sum_{\substack{\{h_1<h_2<\cdots<h_j\}\bigcup\{k_{j+1}<k_{j+2}<\cdots<k_m\}\in P(m)\\ \{h_1<h_2<\cdots<h_j\}\bigcap\{k_{j+1}<k_{j+2}<\cdots<k_m\}=\emptyset}} \\ &\quad \sum_{\substack{i_{h_1}=1,\ldots,n\\ i_{h_2}=1,\ldots,n\\ \vdots\\ i_{h_j}=1,\ldots,n}} \frac{\partial^j}{\partial x_{i_{h_1}}\cdots\partial x_{i_{h_j}}} p(x) V^{h_1}_{i_{h_1}}\cdots V^{h_j}_{i_{h_m}} \|g^{(m-j)}\| \\ &\leq \sum_{j=0}^{m} \sum_{\substack{\{h_1<\cdots<h_j\}\bigcup\{k_{j+1}<\cdots<k_m\}\in P(m)\\ \{h_1<\cdots<h_j\}\bigcap\{k_{j+1}<\cdots<k_m\}=\emptyset}} \|p^{(j)}\|\|g^{(m-j)}\| \\ &= \sum_{j=0}^{m} \begin{pmatrix} m \\ j \end{pmatrix} \|p^{(j)}\|\|g^{(m-j)}\| \end{aligned}$$

which is the desired inequality for the norms above.)

Note that $p^{(i)} \equiv 0, \forall i > m$, so in the summation above we will always have at most $(m+1)$ terms, and, obviously, $\max_i\{\|p^{(i)}\|\} < \infty$ (since p is a polynomial and $\deg(p) \leq m$).

To get a grip on $\|(1/f)^{(i)}\|$ proceed as follows. Introduce

$$g(x) := 1/f(x)$$

and write

$$g(x)f(x) = 1$$

Differentiating the last expression by parts and again using the definition of the norm for a multilinear operator we get

$$\|g^{(k)}\| \leq |1/f^{(0)}| \sum_{i=1}^{k} \begin{pmatrix} k \\ i \end{pmatrix} \|g^{(k-i)}\|\|f^{(i)}\|$$

so the bound on $\|g^{(k)}\| \equiv \|(1/f)^{(k)}\|$ can be computed recursively. Note that for a

more general f at a point z (not just the polynomial) even if

$$\max(\sup_i\{\|f^{(i)}\|\|_z\},1) \leq M < \infty$$

and

$$\max(|1/f(z)|,1) \leq N$$

the expression above will give us

$$\|(1/f)^{(k)}\| \leq 2^{\left(\frac{k(k+1)}{2}-1\right)} N^{k+1} M^k$$

which grows faster than $(k!)$. In our case, where f is a polynomial with $\deg(f) \leq m$, we have at most m terms in the summation. Obviously, $\max(\max_i\{\|f^{(i)}\|\|_z\},1) =: M < \infty$, so the bound

$$\|(1/f)^{(k)}\| \leq 2^{\left(\frac{k(k+1)}{2}-1\right)} N^{k+1} M^k$$

will hold for any $k \leq m$ (or even up to $(m+1)$). On the other hand, for any $k > m$ we can write

$$\|(1/f)^{(k)}\| \leq (NMm)^k \widetilde{L}(k!)$$

where $\widetilde{L} > 0$ is chosen to satisfy

$$\|(1/f)^{(j)}\| \leq (NMm)^j \widetilde{L}(j!), \ \ \forall j \leq m$$

(since if it is true for $1, 2, \ldots, k$ with $k \geq m$, then for $(k+1)$ we can write

$$\begin{aligned}
\|g^{(k+1)}\| &\leq |1/f^{(0)}| \sum_{i=1}^{m} \binom{k}{i} \|g^{(k-i)}\| \|f^{(i)}\| \\
&\leq N((k+1)\|g^{(k)}\|M + (k+1)(k)\|g^{(k-1)}\|M + \cdots \\
&+ \ (k+1)\cdots(k-m+1)\|g^{(k-m)}\|M) \\
&\leq NMm(k+1)(NMm)^k(k!)\widetilde{L}
\end{aligned}$$

and thus is true by induction). Therefore, at a point z such that $f(z) \neq 0$, $\|g^{(k)}\|$ grows at most geometrically in k times $(k!)$.

Coming back to the derivatives of $q(x) = p(x)/f(x)$ at z ($f(z) \neq 0$), redefining

$$M := \max\{\max_i\{\|f^{(i)}\|_z\}, \max_i\{\|p^{(i)}\|_z\}, 1\}$$

and letting

$$N := \max(|1/f(z)|, 1)$$

now we can write

$$\begin{aligned}
\|q^{(k)}\| &\leq \sum_{i=0}^{m} \binom{k}{i} \|p^{(i)}\| \left\| \left(\tfrac{1}{f}\right)^{(k-i)} \right\| \\
&\leq \Big[((NMm)^k \widetilde{L}(k!))M + k((NMm)^{k-1}\widetilde{L}((k-1)!))M \\
&\quad + \cdots + k(k-1)\cdots(k-m+1)((NMm)^{k-m}\widetilde{L}((k-m)!))M\Big] \\
&\leq (m+1)M(NMm)^k \widetilde{L}(k!)
\end{aligned} \tag{3.3.1.17}$$

Therefore, $\|q^{(k)}\|$ (as well as $\|(1/f)^{(k)}\|$) grows no faster then the geometric series in k multiplied by $(k!)$ (as a particular consequence of this we observe that $q(x)$ is real analytic in some neighborhood of z).

Remark 3.3.8 (On the norm of derivatives of $E_k(x)$ for $x \geq 0$). In our setting we are concerned with

$$q(x) = \frac{E_k(x)}{E_{k-1}(x)}$$

Turns out that for the elementary symmetric polynomial $E_j(x)$ at a point $x \in \mathbb{R}^n_+$, there is a connection between its (higher) hyperbolic derivative polynomials $E_j^{(h)}(\cdot)$ evaluated at x and $\| \left(E_j(\cdot)\right)^{(h)} \|_x$. For convenience we introduce $E_0(x) \equiv 1$. Recall

$$E_n^{(n-j)}(x) = (n-j)!E_j(x)$$

and

$$E_j'(x) = \mathbf{1}\nabla_x E_j(x) = \sum_{i=1}^{n} \frac{\partial}{\partial x_i} E_j(x)$$

so that we can write

$$\begin{aligned} \| \left(E_j(x)\right)^{(h)} \|_\infty &= \max_{u_1,u_2,\ldots,u_h,\|u_i\|_\infty \le 1} \left(E_j(x)\right)^{(h)} (u_1, u_2, \ldots, u_h) \\ &= \left(E_j(x)\right)^{(h)} (\mathbf{1}, \ldots, \mathbf{1}) = E_j^{(h)}(x) \end{aligned}$$

since $x \ge 0$. Therefore

$$\| \left(E_j(x)\right)^{(h)} \|_\infty = \left(\frac{1}{(n-j)!} E_n^j(x) \right)^{(h)} = \frac{(n-j+h)!}{(n-j)!} E_{j-h}(x)$$

Note that since $\|u\|_2 \le \|u\|_\infty$ we have

$$\| \left(E_j(x)\right)^{(h)} \| \le \| \left(E_j(x)\right)^{(h)} \|_\infty$$

Considering

$$q(x) = \frac{E_{m+1}(x)}{E_m(x)}$$

(fix $k = m+1$) at some point $x \ge 0$ such that $E_m(x) \ne 0$ we can take

$$M = \max \left\{ \max_{0 \le i \le m+1} \left\{ \frac{(n-(m+1)+i)!}{(n-(m+1))!} E_{m+1-i}(x) \right\}, 1 \right\}$$

and

$$N = \max(|1/E_m(x)|, 1)$$

and $\widetilde{L} \ge N$ to satisfy

$$2^{\left(\frac{j(j+1)}{2}-1\right)} N^{j+1} M^j \le (NMm)^j \widetilde{L}(j!), \;\; 1 \le j \le m$$

e.g., take

$$\widetilde{L} = \max \left\{ \max_{1 \le j \le m} \left\{ \frac{2^{\left(\frac{j(j+1)}{2}-1\right)} N}{m^j (j!)} \right\}, N \right\}$$

to get

$$\|q^{(k)}\|_x \le (m+2) M (NMm)^k \widetilde{L} k! \tag{3.3.1.18}$$

(compare with 3.3.1.17).

The error analysis for Newton iterates starting at x^* (i.e., at $y = 0$)

The vector $x(d)$ can be characterized as the solution to

$$\begin{pmatrix} \text{proj}_{\text{null}(\mathbf{1}^T)} \nabla_y q_d([x_M(y); y]) \\ q_d([x_M(y); y]) \end{pmatrix} = 0$$

which equivalently can be written as

$$g_e(y) := \begin{pmatrix} B\nabla_y q_d([x_M(y); y]) \\ q_d([x_M(y); y]) \end{pmatrix} = 0$$

where

$$B := \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ 1 & 0 & -1 & 0 & \dots & 0 \\ 1 & 0 & 0 & -1 & \dots & 0 \\ & & \dots & & & \\ 1 & 0 & 0 & 0 & \dots & -1 \end{bmatrix}$$

(Note that there is no essential difference between $g(e; y)$ defined earlier and the currently defined $g_e(y)$ in terms of the Newton's iterates, since these two functions differ only by a linear transformation).

We can differentiate $g_e(y)$ with respect to y to obtain

$$D_y g_e(y) = g_e'(y) = \begin{pmatrix} (B\nabla_{yy}^2 q_d) \\ \nabla_y q_d^T \end{pmatrix}$$

We want to compute the inverse of $g_e'(y)$ (if it exists) at $y = 0$.

Letting $A = \nabla_{yy}^2 q_d$ and $w = \nabla_y q_d$ (at $y = 0$), then

$$A^{-1} = [e]\left(-I + \tfrac{\mathbf{1}\mathbf{1}^T}{1+(n-m)}\right)[e]$$

$$z^T := w^T A^{-1} = \mathbf{1}^T \left(-I + \tfrac{\mathbf{1}\mathbf{1}^T}{1+(n-m)}\right)[e] = \left(-\mathbf{1} + \tfrac{(n-m)\mathbf{1}}{1+(n-m)}\right)^T [e] = \left(\tfrac{-1}{1+(n-m)}\right) e^T$$

giving us

$$g'_e(0)^{-1} = A^{-1}\left(\begin{bmatrix} T^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{(\mathbf{1}^T z)}\begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}(z_2, z_3, z_4, \ldots, z_{k-1}, -\mathbf{1}^T z_{-k}, 1)\right)$$

where $k = (n - m)$ and

$$T^{-1} = \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 & 1 \\ -1 & 0 & 0 & \ldots & 0 & 1 \\ 0 & -1 & 0 & \ldots & 0 & 1 \\ & & \ldots & & & \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix}$$

(see Appendix A.2). Obviously, the inverse is defined for all e under consideration (since $e \in \mathbb{R}^{(n-m)}_{++}$).

Given a hyperbolicity direction $d_0 \in \mathbb{L}$, that is, $d_0 = [d_M(e_0); e_0]$ with $e_0 = t\mathbf{1}$ for some $t > 0$, $g_{e_0}(y)$ is a corresponding real analytic function with respect to y in some neighborhood of $y = 0$. Therefore (at $y = 0$)

$$\gamma_0 := \sup_{k \geq 2} \left\| \frac{g'_{e_0}(0)^{-1} g^{(k)}_{e_0}(0)}{k!} \right\|^{\frac{1}{k-1}}$$

must be finite (see [19]).

Define

$$\gamma_e := \sup_{k \geq 2} \left\| \frac{g'_e(0)^{-1} g^{(k)}_e(0)}{k!} \right\|^{\frac{1}{k-1}}$$

(as in [19]; if the reader is not particularly familiar with these authors' complexity analysis for Newton's iterates, look at Appendix B where the main results we need are quoted, or look at the source [19] for their complete exposition).

Recall, in Theorem 3.3.1 we defined the (e_0, ϵ)-*wedge* of a central line $\mathbb{L}_{M^c}$ (or $\mathbb{L}$) as follows: given $e_0 = \tau\mathbf{1}$ ($d_0 \in \mathbb{L}$), $\epsilon > 0$

$$W_{e_0,\epsilon} = \{e \in \mathbb{R}^{n-m} : e = t(e_0 + \delta e), 0 \leq t \leq 1, \delta e \in \text{null}(\mathbf{1}^T), \|\delta e\| \leq \epsilon\}$$

We also introduce (a flat disk)

$$\omega_{e_0,\epsilon} := \{e \in \mathbb{R}^{n-m} : e = e_0 + \delta e, \delta e \in \text{null}(\mathbf{1}^T), \|\delta e\| \leq \epsilon\}$$

(note that $W_{e_0,\epsilon}$ is a convex combination of $\omega_{e_0,\epsilon}$ and 0).

Lemma 3.3.9. *$e_0 \neq 0$ and $\epsilon > 0$ can be chosen such that*

$$\gamma_e \leq \frac{1}{t}\widehat{\gamma}$$

for any $e \in t\omega_{e_0,\epsilon}, 0 < t \leq 1$ (i.e., $e \in W_{e_0,\epsilon}$), for some finite $\widehat{\gamma} > 0$.

Proof. Firstly we will show that the disk $\omega_{e_0,\epsilon}$ can be chosen such that $\forall e \in \omega_{e_0,\epsilon}$, $g_{te}^{(k)}(0)(u_1, u_2, \ldots, u_k)$ will scale component-wise at most proportionally to $(1/t)^{k+1}$ for its first $(n - m - 1)$ components, and to $(1/t)^k$ for its last component (for a fixed arbitrary k-tuple $(u_1, u_2, \ldots, u_k) \in \mathbb{R}^{(n-m)^k}$). Then we will demonstrate that $g'_{te}(0)^{-1}$ will "undo" this scaling (due to its respective components being scaled by t^2 and t), giving us that $g'_e(0)^{-1} g_{te}^{(k)}(0)(u_1, u_2, \ldots, u_k)$ scales at most proportionally to t^{k-1} (for $k \geq 2$). Combined with the analysis of the higher derivatives of $q(x)$ above this will establish the desired result.

For

$$g_e(y) = \begin{pmatrix} B\nabla_y q_d([x_M(y); y]) \\ q_d([x_M(y); y]) \end{pmatrix}$$

we have

$$g_e^{(k)}|_y(u_1,u_2,\ldots,u_k) =$$
$$\begin{pmatrix} B[\widetilde{A}^T I]\left[\frac{1}{d}\right](\nabla_x q(x))^{(k)}|_{\frac{x(y)}{d}}\left(\left[\frac{1}{d}\right]\begin{pmatrix}\widetilde{A}\\ I\end{pmatrix}u_1,\ldots,\left[\frac{1}{d}\right]\begin{pmatrix}\widetilde{A}\\ I\end{pmatrix}u_k\right) \\ (q(x))^{(k)}|_{\frac{x(y)}{d}}\left(\left[\frac{1}{d}\right]\begin{pmatrix}\widetilde{A}\\ I\end{pmatrix}u_1,\ldots,\left[\frac{1}{d}\right]\begin{pmatrix}\widetilde{A}\\ I\end{pmatrix}u_k\right)\end{pmatrix}$$

where for brevity we write $x(y) = [x_M(y); y]$ (and $d = [d_M(e); e]$ as before). This can be written as

$$g_e^{(k)}|_y(u_1,u_2,\ldots,u_k) = \begin{pmatrix} BV^{(k)} \\ (q(x))^{(k)}|_{\frac{x(y)}{d}}\left(\left[\frac{1}{d}\right]\begin{pmatrix}\widetilde{A}\\ I\end{pmatrix}u_1,\ldots,\left[\frac{1}{d}\right]\begin{pmatrix}\widetilde{A}\\ I\end{pmatrix}u_k\right)\end{pmatrix}$$

where $V^{(k)} \in \mathbb{R}^n \cong (\mathbb{R}^n)^*$ (isomorphic to its dual space of continuous linear functionals on $\mathbb{R}^n$) and the corresponding to $V^{(k)}$ unique element in this dual space, $(V^{(k)})^*$, satisfies

$$(V^{(k)})^*(\cdot) = (q(x))^{(k+1)}|_{\frac{x(y)}{d}}\left(\left[\frac{1}{d}\right]\begin{pmatrix}\widetilde{A}\\ I\end{pmatrix}u_1,\ldots,\left[\frac{1}{d}\right]\begin{pmatrix}\widetilde{A}\\ I\end{pmatrix}u_k,\left[\frac{1}{d}\right]\begin{pmatrix}\widetilde{A}\\ I\end{pmatrix}(\cdot)\right)$$

where

$$(V^{(k)})^* : y \mapsto \langle y, V^{(k)}\rangle = y^T V^{(k)}$$

(remember that $V^{(k)}$ depends on $u_1, u_2, \ldots, u_k$). Since for $\mathbb{R}^n$ the norm of $V^{(k)}$ and the (dual) norm of $(V^{(k)})^*$ coincide, we can write

$$\|V^{(k)}\| = \left\|(q(x))^{(k+1)}|_{\frac{x(y)}{d}}\left(\left[\frac{1}{d}\right]\begin{pmatrix}\widetilde{A}\\ I\end{pmatrix}u_1,\ldots,\left[\frac{1}{d}\right]\begin{pmatrix}\widetilde{A}\\ I\end{pmatrix}u_k,\left[\frac{1}{d}\right]\begin{pmatrix}\widetilde{A}\\ I\end{pmatrix}(\cdot)\right)\right\|$$

Moreover, if $\|u_i\| \le 1, \forall i$, we have

$$\|V^{(k)}\| \le \left\|(q(x))^{(k+1)}|_{\frac{x(y)}{d}}\right\|\left\|\left[\frac{1}{d}\right]\right\|^{k+1}\left\|\begin{pmatrix}\widetilde{A}\\ I\end{pmatrix}\right\|^{k+1}$$

where the norms used for two matrices are the operator norms, i.e.,

$$\left\| \left[\frac{1}{d}\right] \right\| = \max_{w \in \mathbb{R}^{n-m}, \|w\| \leq 1} \left\| \left[\frac{1}{d}\right] w \right\| = \max_i \{1/d_i\}$$

and

$$\left\| \begin{pmatrix} \widetilde{A} \\ I \end{pmatrix} \right\| = \max_{w \in \mathbb{R}^{n-m}, \|w\| \leq 1} \left\| \begin{pmatrix} \widetilde{A} \\ I \end{pmatrix} w \right\| \leq \left\| \begin{pmatrix} \widetilde{A} \\ I \end{pmatrix} \right\|_F$$

(Recall the Frobenius norm of a matrix A is defined as $\|A\|_F = \sqrt{\sum_{i,j} (A)_{i,j}^2}$.)

Now, consider a closed ball $B(0, r)$ centered at $e = 0$ of (small) radius $r > 0$ such that

$$d_M(e) \geq \frac{x_M^*}{2}$$

for all $e \in B(0, r)$. Then

$$\left\| \left[\frac{1}{d}\right] \right\| \leq \max \left(\frac{2}{\min_{1 \leq i \leq m} \{(x_M^*)_i\}}, \frac{1}{\min_{1 \leq j \leq (n-m)} \{e_j\}} \right)$$

$\forall e \in B(0, r)$, and if we choose (e_0, ϵ) such that $\epsilon < \|e_0\|$ and $\omega_{e_0,\epsilon} \subseteq B(0, r)$ (i.e., the wedge $W_{e_0,\epsilon} \subset \mathbb{R}_{++}^{(n-m)}$), then, obviously,

$$\left\| \left[\frac{1}{d}\right] \right\| \leq \max \left(\frac{2}{\min_{1 \leq i \leq m} \{(x_M^*)_i\}}, \frac{\sqrt{n-m}}{\|e_0\| - \epsilon} \right)$$

$\forall e \in \omega_{e_0,\epsilon}$, and is finite. Note that if we scale $\omega_{e_0,\epsilon}$ by $t, 0 < t \leq 1$, we have

$$\left\| \left[\frac{1}{d}\right] \right\| \leq \max \left(\frac{2}{\min_{1 \leq i \leq m} \{(x_M^*)_i\}}, \frac{1}{t} \frac{\sqrt{n-m}}{\|e_0\| - \epsilon} \right), \forall e \in t\omega_{e_0,\epsilon}$$

and thus we can write

$$\left\| \left[\frac{1}{d}\right] \right\| \leq \frac{K}{t}$$

(for some $K > 0$) for any $e \in t\omega_{e_0,\epsilon}, 0 < t \leq 1$.

From the analysis of the higher derivatives of $q(x)$ it follows that the pair (e_0, ϵ) can be chosen (small enough) such that

$$\left\| (q(x))^{(k+1)}|_{\frac{x^*}{d}} \right\| < (k+1)! M^{k+1}$$

(for some $M > 0$) for all $e \in W_{e_0,\epsilon}$ (e.g., look at a small enough ball $B(0,r)$ around $e = 0$ intersected with the wedge $W_{e_0,\epsilon}$, such that the denominator in $q(x)|_{x^*/d}$, $E_m(x)|_{x^*/d}$, does not vanish on this ball; all the elementary symmetric polynomials that give rise to the estimate 3.3.1.18 for $\|(q(x))^{(k+1)}\|$ will have finite bounds on their values as continuous functions on a compact as long as d does not cross 0 componentwise, e.g., for $r > 0$ small enough).

Therefore, for $u_1, u_2, \ldots, u_k$, such that $\|u_i\| \leq 1, \forall i$, we can write

$$\|V^{(k)}\| \leq (k+1)! M^{k+1} \left(\frac{K}{t}\right)^{k+1} \left\| \begin{pmatrix} \widetilde{A} \\ I \end{pmatrix} \right\|_F^{k+1}$$

and similarly

$$\left| (q(x))^{(k)}|_{\frac{x^*}{d}} \left(\left[\frac{1}{d}\right] \begin{pmatrix} \widetilde{A} \\ I \end{pmatrix} u_1, \ldots, \left[\frac{1}{d}\right] \begin{pmatrix} \widetilde{A} \\ I \end{pmatrix} u_k \right) \right| \leq k! M^k \left(\frac{K}{t}\right)^k \left\| \begin{pmatrix} \widetilde{A} \\ I \end{pmatrix} \right\|_F^k$$

$\forall e \in t\omega_{e_0,\epsilon}, 0 < t \leq 1$, with properly chosen e_0, ϵ (small enough).

Recall

$$g'_e(0)^{-1} = [e] \left(-I + \frac{\mathbf{1}\mathbf{1}^T}{1 + (n-m)} \right) [e] \begin{pmatrix} & & \\ & & \\ & & \end{pmatrix} \left(\begin{bmatrix} T^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{(\mathbf{1}^T z)} \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} (z_2, z_3, z_4, \ldots, z_{(n-m)-1}, -\mathbf{1}^T z_{-(n-m)}, 1) \right)$$

with

$$z^T = \left(\frac{-1}{1 + (n-m)} \right) e^T$$

For a given $e \in \omega_{e_0,\epsilon}$, consider $\widetilde{e} := te$, for some $0 \leq t \leq 1$. Then

$$\begin{aligned} g'_{\widetilde{e}}(0)^{-1} &= [te]\left(-I + \frac{\mathbf{1}\mathbf{1}^T}{1+(n-m)}\right)[te] \\ &\quad \left(\begin{bmatrix} T^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{t(\mathbf{1}^T z)} \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} (tz_2, tz_3, \ldots, tz_{(n-m)-1}, -t\mathbf{1}^T z_{-(n-m)}, 1)\right) \\ &= t^2[e]\left(-I + \frac{\mathbf{1}\mathbf{1}^T}{1+(n-m)}\right)[e] \\ &\quad \left(\begin{bmatrix} T^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{(\mathbf{1}^T z)} \begin{bmatrix} z_2 & z_3 & \ldots & z_{(n-m)-1} & -\mathbf{1}^T z_{-(n-m)} & 1/t \\ z_2 & z_3 & \ldots & z_{(n-m)-1} & -\mathbf{1}^T z_{-(n-m)} & 1/t \\ & & \vdots & & & \\ z_2 & z_3 & \ldots & z_{(n-m)-1} & -\mathbf{1}^T z_{-(n-m)} & 1/t \end{bmatrix}\right) \end{aligned}$$

so the Jacobian inverse for $g_e(0)$ will scale proportionally to t^2 except for its last column, which will scale proportionally to t. So (in the properly chosen wedge $W_{e_0,\epsilon}$) we can bound the inverse of the Jacobian componentwise (in absolute value) as follows:

$$\left|g'_e(0)^{-1}\right| \leq \left[\begin{array}{c|c} H(\mathbf{1}\mathbf{1}^T_{-(n-m)})t^2 & J\mathbf{1}t \end{array}\right]$$

for some constants $H, J > 0$.

Finally, recalling the definition

$$\gamma_e = \sup_{k\geq 2} \left\| \frac{g'_e(0)^{-1} g_e^{(k)}(0)}{k!} \right\|^{\frac{1}{k-1}}$$

(where, again, the norm is the operator norm) and combining geometric terms in the estimate for $\|g_e^{(k)}(0)\|$ into one (with some constant $\widehat{M}$), we conclude that

$$\gamma_e \leq \sup_{k\geq 2} \left\| \frac{1}{k!} \left[\begin{array}{c|c} H(\mathbf{1}\mathbf{1}^T_{-(n-m)})t^2 & J\mathbf{1}t \end{array}\right] \begin{pmatrix} \mathbf{1}(k+1)!\left(\frac{\widehat{M}}{t}\right)^{k+1} \\ k!\left(\frac{\widehat{M}}{t}\right)^{k} \end{pmatrix} \right\|^{\frac{1}{k-1}} \leq \frac{\widehat{\gamma}}{t}$$

$\forall e \in t\omega_{e_0,\epsilon}, 0 < t \leq 1$ for some $0 < \widehat{\gamma} < \infty$ (the extra term in the factorial does not cause any troubles since $(k+1)^{1/(k-1)}$ is bounded by 3 for $k \geq 2$). $\square$

Theorem 3.3.10. *(e_0, ϵ) can be chosen such that there exists $\Psi > 0$ where for $\forall e \in W_{e_0,\epsilon}, e \neq 0$, the difference between the true optimal solution $y(e)$ and the Newton step $N_{g_e}(0)$ for $g_e(y) = 0$ is bounded by*

$$\|y(e) - N_{g_e}(0)\| \leq \|N_{g_e}(0)\|^2 \frac{\Psi}{(\mathbf{1}^T e)}$$

(that is, has a "quadratic error term" as compared to the Newton step itself).

Proof. We start by considering the set

$$\omega_{e_0,\epsilon} = \{e \in \mathbb{R}^{n-m} : e = e_0 + \delta e, \delta e \in \text{null}(\mathbf{1}^T), \|\delta e\| \leq \epsilon\}$$

with $\epsilon < \|e_0\|$ (recall $W_{e_0,\epsilon}$ is a convex combination of $\omega_{e_0,\epsilon}$ and the origin $e = 0$).

Let $\gamma_e \leq \frac{1}{t}\widehat{\gamma}$ for $\forall e \in t\omega_{e_0,\epsilon}, 0 \leq t \leq 1$ (as in Lemma 3.3.9). That is, if we introduce

$$\widehat{\gamma}_t := \sup_{e \in t\omega_{e_0,\epsilon}} \gamma(g_e, 0)$$

(for some fixed $0 < t \leq 1$), then

$$\widehat{\gamma}_t = \frac{1}{t}\widehat{\gamma}$$

For any positive constant α_0, the pair (e_0, ϵ) can be chosen such that

$$\widehat{\alpha} = \widehat{\gamma}\widehat{\beta} < \alpha_0$$

with

$$\widehat{\beta} = \sup_{e \in \omega_{e_0,\epsilon}} \|NS_{g_e}(0)\|$$

(recall computing $NS_g(e, y)$ at $y = 0$). In particular, we can choose α_0 as coming from Smale's convergence analysis for Newton iterates (see Appendix B), e.g.,

$\widehat{\alpha} = .01 < .03 < \alpha_0$. The condition $\widehat{\alpha} < \alpha_0$ insures that $y_0 = 0$ is an "approximate zero" for $g_e(y)$ (in the terminology of Appendix B). In particular, with this choice of (e_0, ϵ) we have

$$\|y(e) - y_0\| \leq 2\|NS_{g_e}(0)\|$$

and moreover, if

$$2\widehat{\beta}\widehat{\gamma} \leq \widehat{u} = .06 < 1 - \frac{\sqrt{2}}{2}$$

it follows that

$$\|y(e) - y_0\|\gamma_e \leq 2\|NS_{g_e}(0)\|\gamma_e \leq 2\widehat{\beta}\widehat{\gamma} \leq \widehat{u} < 1 - \frac{\sqrt{2}}{2}$$

and thus

$$\gamma(g_e, y(e)) \leq \frac{\gamma(g_e, 0)}{\psi(\widehat{u})(1 - \widehat{u})}$$

(and more generally for all y in the ball centered at $y_0 = 0$ of radius $2\|NS_{g_e}(0)\|$; this clarifies the first part of the Remark 3.3.3 precisely) so that for this particular choice of (e_0, ϵ) (with $\widehat{u} \leq .06$)

$$\gamma(g_e, y(e)) \leq \frac{685}{494}\gamma(g_e, 0)$$

At this point we will complete the clarification of the Remark 3.3.3. We are interested in the solution to $g_e(y) = 0$ satisfying $\nabla_y q_d([x_M(y); y]) = \tau \mathbf{1}$ for some $\tau > 0$. Recall that the value of τ corresponding to the first Newton's iterate, $N_{g_e}(0) = NS_{g_e}(0) = NS_g(e, 0)$, for solving $g_e(y) = 0$ (recall that this is the same as for solving $g(e; y) = 0$ w.r.t. y) starting at $y_0 = 0$ is $\tau_0 := \frac{(n-m)}{\mathbf{1}^T e}$. If instead of $g(e; y) = 0$ (or $g_e(y) = 0$) we consider

$$\widetilde{g}(e; (y, \tau)) := \begin{pmatrix} \nabla_y q_d([x_M(y); y]) - \tau \mathbf{1} \\ q_d([x_M(y); y]) \end{pmatrix} = 0$$

then this equation will give us precisely the same Newton's iterates (w.r.t. (y,τ)) in y as $g(e;y)=0$ (or $g_e(y)=0$) and precisely the same constant $\gamma(\widehat{g}(e;(y,\tau),(y_0,\tau_0))$ (same as $\gamma(g(e;y),y_0)=\gamma(g_e(y),y_0)$; in terms of the notation in Appendix B) at $y_0=0$, since the "free linear term", $\tau\mathbf{1}$, will have no effect on the higher derivatives of $\widetilde{g}(e;(y,\tau))$ (w.r.t. (y,τ)). Now, is for $\widetilde{g}(e;(y,\tau))=0$ we start the Newton's method at a point $(y,\tau)_0=(0,\tau_0)$, then for the choice of (e_0,ϵ) as above, the associated root of this equation, (y^*,τ^*), will be contained in the ball centered at $(0,\tau_0)$ of the same radius $2\|NS_{g_e}(0)\|$ not exceeding $2\widehat{\beta}$ (note that no change in τ is needed for the first iterate). In particular, we get that $|\tau^*-\tau_0|\leq 2\widehat{\beta}$ as well. So, as long as e pick (e_0,ϵ) such that

$$\tau_0-2\widehat{\beta}>0$$

the corresponding solution to $\widehat{g}(e;(y,\tau))=0$ will have $\tau^*>0$. Indeed, this inequality can be easily satisfied by choosing e_0 and ϵ small enough (recall the expression for $NS_{g_e}(0)$).

Now, back to the proof. We can write

$$\begin{aligned}\|N_{g_e}(0)-y(e)\| &\leq \frac{\gamma(g_e,y(e))\|y_0-y(e)\|^2}{\psi(u)}\\ &\leq \frac{685}{494}\widehat{\gamma}\frac{\|y_0-y(e)\|^2}{\psi(u)}\end{aligned}$$

where

$$\begin{aligned}u &= \|y_0-y(e)\|\gamma(g_e,y(e))\leq 2\|NS_{g_e}(0)\|\frac{685}{494}\widehat{\gamma}\\ &\leq 2\frac{685}{494}\widehat{\gamma}\widehat{\beta}\leq\widehat{u}\frac{685}{494}=0.06\left(\frac{685}{494}\right)\end{aligned}$$

and therefore

$$\begin{aligned}\|N_{g_e}(0)-y(e)\| &\leq \left(\frac{685}{494}\right)\left(\frac{533}{363}\right)\widehat{\gamma}\|y_0-y(e)\|^2\\ &\leq 3\widehat{\gamma}\|y_0-y(e)\|^2\leq 12\widehat{\gamma}\|NS_{g_e}(0)\|^2\end{aligned}$$

Given a pair (e_0, ϵ) (thus defining $\omega_{e_0,\epsilon}$), consider how $\widehat{\beta}$ changes if we consider $t\omega_{e_0,\epsilon}$ for some $0 \leq t \leq 1$. Recall that (see 3.3.1.14)

$$NS_{g_e}(0) = \alpha\left(-e_\perp - \frac{e_\perp^2}{e_\parallel}\right)$$

where $e = e_\perp + e_\parallel$ with $e_\parallel = t\mathbf{1}$ and where

$$\alpha = \alpha(e) = \frac{E_m}{E_{m-1}}\left(\frac{x_M^*}{d_M(e)}\right)$$

satisfies

$$\left|\alpha(e) - \frac{1}{m}\right| \leq K\|e\|$$

Since $\alpha(e)$ is Lipschitz continuous at 0, the pair (e_0, ϵ) can be chosen such that

$$\frac{1}{2m} \leq \alpha(e) \leq \frac{3}{2m}, \forall e \in W_{e_0,\epsilon}$$

Now if we consider $\widetilde{e} = te$ for some $e \in \omega_{e_0,\epsilon}$ we have

$$\begin{aligned}
\|NS_{g_{\widetilde{e}}}(0)\| &= \left\|\alpha(\widetilde{e})\left(-\widetilde{e}_\perp - \tfrac{\widetilde{e}_\perp^2}{\widetilde{e}_\parallel}\right)\right\| \\
&= |\alpha(\widetilde{e})|\left\|\left(-\widetilde{e}_\perp - \tfrac{\widetilde{e}_\perp^2}{\widetilde{e}_\parallel}\right)\right\| \\
&= |\alpha(\widetilde{e})|\left\|\left(-te_\perp - \tfrac{(te_\perp)^2}{te_\parallel}\right)\right\| \\
&\leq t\tfrac{3}{2m}\left\|\left(-e_\perp - \tfrac{e_\perp^2}{e_\parallel}\right)\right\|
\end{aligned}$$

and combined with

$$\|NS_{g_e}(0)\| = \left\|\alpha(e)\left(-e_\perp - \frac{e_\perp^2}{e_\parallel}\right)\right\| \geq \frac{1}{2m}\left\|\left(-e_\perp - \frac{e_\perp^2}{e_\parallel}\right)\right\|$$

this gives us

$$\|NS_{g_{\widetilde{e}}}\| \leq 3t\|NS_{g_e}\|$$

Thus, if we define

$$\widehat{\beta}_t = \sup_{e \in t\omega_{e_0,\epsilon}} |\alpha(e)|\left\|\left(-e_\perp - \frac{e_\perp^2}{e_\parallel}\right)\right\|$$

for $0 \leq t \leq 1$, then

$$\widehat{\beta}_t \leq 3t\widehat{\beta}$$

As a consequence of the bounds above we have

$$\widehat{\beta}_t\widehat{\gamma}_t \leq .03 < \alpha_0$$

and

$$2\widehat{\beta}_t\widehat{\gamma}_t \leq .06 = \widehat{u}$$

and therefore the analysis of the Newton step iterates above will apply for any point in $W_{e_0,\epsilon}$, resulting in the following bound for $\widetilde{e} \in W_{e_0,\epsilon}$ ($\widetilde{e} = te$, $e \in \omega_{e_0,\epsilon}$, $0 < t \leq 1$)

$$\begin{aligned} \|N_{g_{\widetilde{e}}}(0) - y(\widetilde{e})\| &\leq \left(\tfrac{685}{494}\right)\left(\tfrac{533}{363}\right)\widehat{\gamma}_t\|y_0 - y(\widetilde{e})\|^2 \\ &\leq \left(\tfrac{685}{494}\right)\left(\tfrac{533}{363}\right)\widehat{\gamma}_t\|2NS_{g_e}(0)\|^2 \\ &\leq 12\widehat{\gamma}_t\|NS_{g_{\widetilde{e}}}(0)\|^2 \\ &= \tfrac{12\widehat{\gamma}}{t}\|NS_{g_{\widetilde{e}}}(0)\|^2 \end{aligned} \tag{3.3.1.19}$$

which is what we wanted to show. □

Second exercise in ODE

We refine our ODE setting analyzed in Lemma 3.3.6 to bring it closer to the dynamics that we consider in Theorem 3.3.1 (we in particular, look at the ODE in $\mathbb{R}^{n-m} \equiv \mathbb{R}^{k+1}$ instead of in $\mathbb{R}^2$). Consider the following system of ODE:

$$\begin{cases} \dot{x}_1 &= -x_1 + (\eta + O_1(|x_1| + \|x_2\|))O_{2,1}\left(\frac{\|x_2\|^2}{x_1}\right) \\ &\quad + O_{3,1}\left(\frac{\|x_2\|^2}{x_1}\right) =: f_1(x_1, x_2, t) \\ \dot{x}_{2,i} &= -x_{2,i} + (\eta + O_1(|x_1| + \|x_2\|))(-x_{2,i}) \\ &\quad + (\eta + O_1(|x_1| + \|x_2\|))O_{2,2,i}\left(\frac{\|x_2\|^2}{x_1}\right) \\ &\quad + O_{3,2,i}\left(\frac{\|x_2\|^2}{x_1}\right) =: f_{2,i}(x_1, x_2, t) \text{ for } i = 1, \ldots, k \end{cases} \tag{3.3.1.20}$$

with some starting point ($t_0 = 0$)

$$x_1(0) = \bar{x}_1, x_2(0) = \bar{x}_2 \tag{3.3.1.21}$$

$\bar{x}_1 = 1, \|\bar{x}_2\|$ being possibly $\ll 1$ and $0 < \eta$ (we denote the solution to 3.3.1.20, 3.3.1.21 as $(x_1(t), x_2(t))$.). Assume, as before, $f_1, f_{2,i}$ are continuous on $\{(x_1, x_2) \in \mathbb{R} \times \mathbb{R}^k : x_1 \geq 0\}$, and thus the solution to this initial value problem exists. Assume $\eta < 1/2$ (see the note in the proof on the choice of η).

We want to understand the behavior of the system as $t \uparrow \infty$ under some additional assumptions on $O_i, O_{i,j,k} : \mathbb{R} \to \mathbb{R}$. In particular, $|O_1(y)| \leq K_1|y|, |O_{i,j}(y)| \leq K_i|y|, |O_{i,j,k}(y)| \leq K_i|y|$. While we impose no assumptions on the magnitude of K_2, K_3, we will suppose that K_1 is relatively small with respect to η, for example $K_1 < \frac{\eta}{6}$ (combined with the assumption on the magnitude of η, namely $\eta < 1/2$; see later discussion in the proof).

As was the case for $\mathbb{R}^2$, intuitively, one can look at system 3.3.1.20 as

$$\dot{x}_1 \approx -x_1$$
$$\dot{x}_2 \approx (-1 - \eta)x_2$$

($x_2 \in \mathbb{R}^k$ now) so one would expect to have the solution of the form $x_1(t) \approx \bar{x}_1 e^{-t}, x_2(t) \approx \bar{x}_2 e^{(-1-\eta)t}$.

Lemma 3.3.11. *Suppose $0 < \eta < 1/2$ and $K_1 < \eta/6$. Then the constant $\epsilon > 0$ can be chosen such that for any starting point $(\bar{x}_1, \bar{x}_2)$ in a wedge*

$$W_\epsilon := \left\{(x_1, x_2) \in \mathbb{R} \times \mathbb{R}^k : 0 \leq x_1 \leq 1, 0 \leq \|x_2\| \leq \epsilon x_1\right\}$$

the solution to 3.3.1.20, 3.3.1.21 will approach the origin, staying in W_ϵ, $\forall t \geq 0$. Moreover, if $\|x_2(0)\| \neq 0$, then as $t \uparrow \infty$

$$x_1(t) \sim C_1 e^{-t}$$
$$\|x_2(t)\| \sim C_2 e^{-(1+\eta)t}$$

for some constants C_1, C_2 (> 0).

Proof. Since we are interested in the dynamics of $\|x_2(t)\|$ rather than how $x_2(t)$ evolves coordinate-wise, we will be choosing the system of (orthogonal) coordinates for x_2 somewhat arbitrarily (as will be illustrated further in the proof) and will have to work our way through the (possibly constantly changing) choice of the coordinate system as the argument evolves. Besides that, the proof is very similar to that of Lemma 3.3.6 but is presented in its fullness since the noted details are substantially different.

Consider the solution to 3.3.1.20, 3.3.1.21 corresponding to the initial point $(\bar{x}_1, \bar{x}_2)$ in a flat disk

$$\omega_\epsilon := \left\{(x_1, x_2) \in \mathbb{R} \times \mathbb{R}^k : x_1 = 1, \|x_2\| \leq \epsilon\right\}$$

For now, without loss of generality, assume $(x_2)_0 > 0$. (If $(x_2)_0 = 0$ the result follows almost immediately: as a direct consequence of Lipschitz continuity we get $x_2(t) \equiv 0$, $\forall t \geq 0$, and therefore $x_1(t) = e^{-t}$.)

Since the solution to 3.3.1.20, 3.3.1.21 is continuous, given some $\alpha, \beta > 1$, we can pick $\Delta t > 0$ small enough such that $x_1(t) \in [1/\alpha, \beta]$, $x_2(t) \geq 0$ on $[0, \Delta t]$. We will refine our choice of β a bit later.

From Lipschitz continuity of O_1 it follows that

$$|O_1(|x_1| + \|x_2\|)| \leq K_1|(|x_1| + \|x_2\|)| \leq K_1(|x_1| + \|x_2\|) \leq K_1\beta + K_1\|x_2\|$$

on $[0, \Delta t]$. At this point our argument becomes more involved than that of Lemma 3.3.6. Choose the coordinate system for x_2 as follows (by possibly rotating the current basis for x_2; note that the ODE 3.3.1.20 is rotation-invariant

with respect to x_2): let

$$\bar{x}_{2,1} = 2\zeta$$
$$\bar{x}_{2,2} = \zeta$$
$$\bar{x}_{2,3} = \zeta$$
$$\vdots$$
$$\bar{x}_{2,k} = \zeta$$

with $\|\bar{x}_2\|$ given, then

$$\|\bar{x}_2\|^2 = 4\zeta^2 + (k-1)\zeta^2$$

so

$$\zeta = \frac{\|\bar{x}_2\|}{\sqrt{3+k}}$$

Now we have

$$\begin{aligned}
f_{2,i}(x_1(0), x_2(0), 0) &= -\bar{x}_{2,i} + (\eta + O_1(|\bar{x}_1| + \|\bar{x}_2\|))(-\bar{x}_{2,i}) \\
&+ (\eta + O_1(|\bar{x}_1| + \|\bar{x}_2\|))O_{2,2,i}\left(\frac{\|\bar{x}_2\|^2}{\bar{x}_1}\right) \\
&+ O_{3,2,i}\left(\frac{\|\bar{x}_2\|^2}{\bar{x}_1}\right) \\
&\leq -\bar{x}_{2,i} - \eta\bar{x}_{2,i} + \bar{x}_{2,i}(1K_1 + K_1\|\bar{x}_2\|) \\
&+ (\eta + 1K_1 + K_1\|\bar{x}_2\|)K_2\left(\frac{\|\bar{x}_2\|^2}{\bar{x}_1}\right) + K_3\left(\frac{\|\bar{x}_2\|^2}{\bar{x}_1}\right) \\
&= \bar{x}_{2,i}(-1 - \eta + K_1 + K_1\|\bar{x}_2\|) \\
&+ \|\bar{x}_2\|^2((\eta + K_1 + K_1\|\bar{x}_2\|)K_2 + K_3) < 0
\end{aligned}$$

$\forall i = 1, \ldots, k$, provided $\|\bar{x}_2\|$ is small enough. Assume we start with such $\bar{x}_2$. Consequently $x_{2,i}(t)$ will decrease in some (possibly even smaller than $[0, \Delta t]$) neighborhood of $t_0 = 0$, say $[0, \widehat{\tau}]$.

We introduce the notion of the "slowest decaying" trajectory for $x_{2,i}(t), t \geq 0$, to make further analysis possible. Let

$$\tau_i := \inf\{t \geq 0 : 0 \leq x_{2,i}(t) \leq x_{2,k}(t) \text{ for some } k \neq i\}$$

be the *first upcrossing time* for $x_{2,i}(t), t \geq 0$. Let

$$\tau = \max_i \min\{\tau_i, \widehat{\tau}\}$$

and

$$j = \arg\max_i \min\{\tau_i, \widehat{\tau}\}$$

be the index of the slowest decaying trajectory $x_{2,j}(t), t \geq 0$ (note that due to how j is defined, $x_{2,j}(t)$ will dominate $x_{2,i}(t), i \neq j$ for $t \in [0, \tau]$). (If the starting point $\bar{x}_2$ is assumed to have all coordinates distinct it immediately follows that $\tau > 0$ by continuity of $x_2(t)$.) Obviously (with this choice of coordinate system for x_2), $j \equiv 1$.

With this in mind, for $t \in [0, \Delta t] \bigcap [0, \tau]$ we can write

$$\begin{aligned}
|f_{2,i}(x_1, x_2, t) - (-1-\eta)x_{2,i}| &\leq (K_1\beta + K_1\|\bar{x}_2\|)x_{2,i} \\
&+ (\eta + \beta K_1 + K_1\|\bar{x}_2\|)K_2\alpha\|x_2\|^2 + K_3\alpha\|x_2\|^2 \\
&\leq (K_1\beta + K_1\|\bar{x}_2\|)x_{2,i} + (\eta + \beta K_1 \\
&+ K_1\|\bar{x}_2\|)K_2\alpha\left(x_{2,j}\sum_i |\bar{x}_{2,i}|\right) \\
&+ K_3\alpha\left(x_{2,j}\sum_i |\bar{x}_{2,i}|\right)
\end{aligned}$$

since

$$\|x_2\|^2 = x_{2,1}^2 + x_{2,2}^2 + \cdots + x_{2,k-1}^2 \leq x_{2,j}\left(\sum_i |x_{2,i}|\right) \leq x_{2,j}\left(\sum_i |\bar{x}_{2,i}|\right)$$

and, in particular,

$$|f_{2,j}(x_1, x_2, t) - (-1-\eta)x_{2,j}| \leq \gamma x_{2,j} \tag{3.3.1.22}$$

where

$$\begin{aligned}\gamma &:= K_1\beta + K_1\|\bar{x}_2\| + (\eta + \beta K_1 + K_1\|\bar{x}_2\|)K_2\alpha\left(\sum_i |\bar{x}_{2,i}|\right) \\ &\quad + K_3\alpha\left(\sum_i |\bar{x}_{2,i}|\right)\end{aligned}$$

Assume $\beta \leq 2$ and $K_1 < \frac{\eta}{6} < \frac{\eta}{2}$. Then by choosing $\|\bar{x}_2\|$ (i.e., ϵ) sufficiently small, we may also assume

$$-1 - \eta + \gamma < 0 \tag{3.3.1.23}$$

This ensures that $x_{2,j}(t)$ is decreasing for all $t \in [0, \min\{\tau, \Delta t\}]$ (by Proposition 3.3.7 as in the proof of Lemma 3.3.6), and we have

$$\bar{x}_{2,j}e^{(-1-\eta-\gamma)t} \leq x_{2,j}(t) \leq \bar{x}_{2,j}e^{(-1-\eta+\gamma)t} \tag{3.3.1.24}$$

on $[0, \Delta t]$.

Now let us consider $x_{2,i}(t)$ for $i \neq j$ (i.e., $i \geq 2$). Replacing $\|x_2\|^2$ on $t \in [0, \min\{\tau, \Delta t\}]$ with

$$\|x_2\|^2 = \sum_{l=1}^{k} x_{2,l}^2 \leq \sum_{l=1}^{k} x_{2,j}^2 \leq \sum_{l=1}^{k} \left(\bar{x}_{2,j}e^{(-1-\eta+\gamma)t}\right)^2 = k(\bar{x}_{2,j}^2 e^{2(-1-\eta+\gamma)t}$$

we write

$$\begin{aligned}|f_{2,i}(x_1, x_2, t) - (-1-\eta)x_{2,i}| &\leq (K_1\beta + K_1\|\bar{x}_2\|)x_{2,i} \\ &+ [(\eta + \beta K_1 + K_1\|\bar{x}_2\|)K_2\alpha + K_3\alpha]\, k\bar{x}_{2,j}^2 e^{2(-1-\eta+\gamma)t}\end{aligned}$$

If we can guarantee that

$$\begin{aligned}0 &> -1 - \eta + \xi \\ &> -1 - \eta - \xi \\ &> 2(-1 - \eta + \gamma)\end{aligned} \tag{3.3.1.25}$$

where

$$\xi := (K_1\beta + K_1\|\bar{x}_2\|) + [(\eta + \beta K_1 + K_1\|\bar{x}_2\|)K_2\alpha + K_3\alpha]\, k\bar{x}_{2,j}^2$$

which can be satisfied by picking sufficiently small $\|\bar{x}_2\|$ (i.e., ϵ), then we also have the following bounds on $x_{2,i}(t), i \geq 2$, for $t \in [0, \min\{\tau, \Delta t\}]$

$$\bar{x}_{2,i}e^{(-1-\eta-\xi)t} \leq x_{2,i}(t) \leq \bar{x}_{2,i}e^{(-1-\eta+\xi)t} \tag{3.3.1.26}$$

Note on the choice of sufficiently small ϵ: observe that to get the inequality 3.3.1.25 achievable with just the choice of $\bar{x}_2$ we need to have constants K_1, η and β satisfying

$$0 > -1 - \eta + 3K_1\beta \quad \text{and} \quad -1 + 2K_1\beta < 0$$

Namely, if $0 < \eta < 1/2$, $\beta \leq 2$, $K_1 < \eta/6 < \eta/2$ (as in the assumptions of the lemma), then the inequalities above can be satisfied by choosing $\|\bar{x}_2\|$ (i.e., ϵ) small enough. In the setting of Theorem 3.3.1 when we apply this result, we will have control over K_1 in particular. (However, $\eta < 1/2$ is a technical assumption that just simplifies some bound derivations we encountered.)

Let us consider $x_1(t)$. From

$$\begin{aligned} f_1(x_1, x_2, t) &= -x_1 + (\eta + O_1(|x_1| + \|x_2\|))O_{2,1}\left(\frac{\|x_2\|^2}{x_1}\right) \\ &+ O_{3,1}\left(\frac{\|x_2\|^2}{x_1}\right) \end{aligned}$$

and the bound 3.3.1.24 on $x_{2,j}(t)$ on $[0, \min\{\tau, \Delta t\}]$ we get

$$|f_1(x_1, x_2, t) + x_1| \leq ((\eta + K_1\beta + K_1\|(x_2)_0\|)\alpha K_2 + \alpha K_3)k\bar{x}_{2,j}^2 e^{2(-1-\eta+\gamma)t}$$

Letting

$$\psi := ((\eta + K_1\beta + K_1\|(x_2)_0\|)\alpha K_2 + \alpha K_3)k\bar{x}_{2,j}^2$$

we claim that if

$$\begin{aligned} 0 &> -1+\psi \\ &> -1-\psi \\ &> 2(-1-\eta+\gamma) \end{aligned} \tag{3.3.1.27}$$

then for $t \in [0, \min\{\tau, \Delta t\}]$:

$$\bar{x}_1 e^{(-1-\psi)t} \leq x_1(t) \leq \bar{x}_1 e^{(-1+\psi)t} \tag{3.3.1.28}$$

Note that the condition 3.3.1.27 is easily met for sufficiently small $\|\bar{x}_2\|$ (recall that $K_1 < \eta/2, \beta \leq 2, 0 < \eta < 1/2$). This also shows that $x_1(t)$ is decreasing. (The bounds follow from Proposition 3.3.7 presented earlier.)

To summarize what we have so far, under the conditions 3.3.1.23, 3.3.1.25, 3.3.1.27 (all of which can be simultaneously satisfied by picking $\bar{x}_2$ below a certain threshold ϵ in norm) we have the following bounds on $x_1(t), x_2(t)$ for $t \in [0, \min\{\tau, \Delta t\}]$:

$$\bar{x}_1 e^{(-1-\psi)t} \leq x_1(t) \leq \bar{x}_1 e^{(-1+\psi)t}$$

$$\bar{x}_{2,j} e^{(-1-\eta-\gamma)t} \leq x_{2,j}(t) \leq \bar{x}_{2,j} e^{(-1-\eta+\gamma)t}$$

$$\bar{x}_{2,i} e^{(-1-\eta-\xi)t} \leq x_{2,i}(t) \leq \bar{x}_{2,i} e^{(-1-\eta+\xi)t}, \forall i \neq j$$

where

$$\gamma = K_1\beta + K_1\|\bar{x}_2\| + ((\eta + \beta K_1 + K_1\|\bar{x}_2\|)K_2 + K_3)\alpha \left(\textstyle\sum_i |\bar{x}_{2,i}|\right)$$

$$\xi = K_1\beta + K_1\|\bar{x}_2\| + [(\eta + \beta K_1 + K_1\|\bar{x}_2\|)K_2\alpha + K_3\alpha]\, k\bar{x}_{2,j}^2$$

$$\psi = ((\eta + K_1\beta + K_1\|\bar{x}_2\|)\alpha K_2 + \alpha K_3)k\bar{x}_{2,j}^2$$

Furthermore, we can pick $\bar{x}_2$ such that $\gamma \geq \xi (> 0)$. With this choice of $\bar{x}_2$ it follows that

$$\bar{x}_{2,i} e^{(-1-\eta-\gamma)t} \leq x_{2,i}(t) \leq \bar{x}_{2,i} e^{(-1-\eta+\gamma)t}, \text{ for } \forall i$$

and therefore

$$\|\bar{x}_2\|e^{(-1-\eta-\gamma)t} \leq \|x_2(t)\| \leq \|\bar{x}_2\|e^{(-1-\eta+\gamma)t}$$

for $t \in [0, \min\{\tau, \Delta t\}]$.

What can be said about $\min\{\tau, \Delta t\}$ (in particular, we are interested if this time intervals can get arbitrarily small as we will be re-scaling our ODE later on in the proof)? With our choice of the coordinate system for x_2

$$\bar{x}_{2,1} = 2\zeta$$

$$\bar{x}_{2,i} = \zeta, \text{ for all } i > 1$$

τ can be bounded from below by

$$\widetilde{\tau} := \sup\{t \geq 0 : (x_{2,1})_0 e^{(-1-\eta-\gamma)t} \geq (x_{2,2})_0 e^{(-1-\eta+\gamma)t}\}$$

that is

$$2\zeta e^{(-1-\eta-\gamma)\widetilde{\tau}} = \zeta e^{(-1-\eta+\gamma)\widetilde{\tau}}$$

$$2 = e^{2\gamma\widetilde{\tau}}$$

$$\widetilde{\tau} = \tfrac{1}{2\gamma}\ln 2$$

and hence

$$\tau \geq \frac{1}{2\gamma}\ln 2$$

In its turn, Δt can be bounded by

$$\widetilde{\Delta t} := \sup\{t \geq 0 : (x_1)_0 e^{(-1-\psi)t} \geq \alpha\}$$

that is

$$e^{(-1-\psi)\widetilde{\Delta t}} = \alpha$$

$$(-1-\psi)\widetilde{\Delta t} = \ln\alpha$$

$$\widetilde{\Delta t} = \tfrac{-1}{(1+\psi)}\ln\alpha$$

and therefore

$$\Delta t \geq \frac{-1}{(1+\psi)}\ln\alpha$$

(These two bounds will prevent us from having an accumulation point for t other than $t \to \infty$ when we iterate this argument.) We denote $T := \min\{\Delta t, \tau\}$.

We have established exponential bounds on $x_1(t), \|x_2(t)\|$ for $t \in [0, T]$. It remains to show that similar bounds will hold uniformly for all $t \geq 0$.

In order to do so consider a re-scaling of 3.3.1.20, 3.3.1.21. Given $(x_1)_T, (x_2)_T$, we want to put 3.3.1.20, 3.3.1.21 to its "original" setting corresponding to a starting point at $t_0 = T$.

Introduce

$$\widetilde{x}_1 := \frac{x_1}{(x_1)_T}, \widetilde{x}_2 := \frac{x_2}{(x_2)_T}\bar{x}_2$$

and let us observe what will happen to $O_i, O_{i,j}, O_{i,j,k}$ under this transformation.

Note that

$$\frac{(\widetilde{x}_2)^2}{\widetilde{x}_1} = \frac{x_2^2}{x_1}(x_1)_T \left(\frac{\bar{x}_2}{(x_2)_T}\right)^2$$

Consider an equivalent system to 3.3.1.20

$$\left\{\begin{array}{rcl} \frac{1}{(x_1)_T}\dot{x}_1 & = & \frac{1}{(x_1)_T}\left(-x_1 + (\eta + O_1(|x_1| + \|x_2\|))O_{2,1}\left(\frac{\|x_2\|^2}{x_1}\right)\right. \\ & + & \left. O_{3,1}\left(\frac{\|x_2\|^2}{x_1}\right)\right) \\ \frac{\bar{x}_{2,i}}{(x_{2,i})_T}\dot{x}_{2,i} & = & \frac{\bar{x}_{2,i}}{(x_{2,i})_T}\left(-x_{2,i} + (\eta + O_1(|x_1| + \|x_2\|))(-x_{2,i})\right. \\ & + & (\eta + O_1(|x_1| + \|x_2\|))O_{2,2,i}\left(\frac{\|x_2\|^2}{x_1}\right) \\ & & \left. + O_{3,2,i}\left(\frac{\|x_2\|^2}{x_1}\right)\right) \text{ for } i = 1, \ldots, k \end{array}\right. \tag{3.3.1.29}$$

This can be rewritten as

$$\left\{\begin{array}{rcl} \dot{\widetilde{x}}_1 & = & -\widetilde{x}_1 + (\eta + O_1(|x_1| + \|x_2\|))O_{2,1}\left(\left\|\frac{\widetilde{x}_2}{\bar{x}_2}(x_2)_T\right\|^2 \frac{1}{\widetilde{x}_1(x_1)_T}\right)\frac{1}{(x_1)_T} \\ & & + O_{3,1}\left(\left\|\frac{\widetilde{x}_2}{\bar{x}_2}(x_2)_T\right\|^2 \frac{1}{\widetilde{x}_1(x_1)_T}\right)\frac{1}{(x_1)_T} \\ \dot{\widetilde{x}}_{2,i} & = & -\widetilde{x}_{2,i} + (\eta + O_1(|x_1| + \|x_2\|))(-\widetilde{x}_{2,i}) \\ & & + (\eta + O_1(|x_1| + \|x_2\|))O_{2,2,i}\left(\left\|\frac{\widetilde{x}_2}{\bar{x}_2}(x_2)_T\right\|^2 \frac{1}{\widetilde{x}_1(x_1)_T}\right)\frac{\bar{x}_{2,i}}{(x_{2,i})_T} \\ & & + O_{3,2,i}\left(\left\|\frac{\widetilde{x}_2}{\bar{x}_2}(x_2)_T\right\|^2 \frac{1}{\widetilde{x}_1(x_1)_T}\right)\frac{\bar{x}_{2,i}}{(x_{2,i})_T} \text{ for } i = 1, \ldots, k \end{array}\right. \tag{3.3.1.30}$$

We assumed that $|O_1(y)| \leq K_1|y|, |O_{i,j}(y)| \leq K_i|y|, |O_{i,j,k}(y)| \leq K_i|y|$, therefore

$$\begin{aligned}
& \left| O_{2,1}\left(\left\| \frac{\widetilde{x}_2}{\bar{x}_2}(x_2)_T \right\|^2 \frac{1}{\widetilde{x}_1(x_1)_T} \right) \frac{1}{(x_1)_T} \right| \\
= & \left| O_{2,1}\left(\sum_{i=1}^{k} \left(\frac{\widetilde{x}_{2,i}}{\bar{x}_{2,i}}(x_{2,i})_T \right)^2 \frac{1}{\widetilde{x}_1(x_1)_T} \right) \frac{1}{(x_1)_T} \right| \\
\leq & K_2 \left| \left(\sum_{i=1}^{k} \left(\frac{\widetilde{x}_{2,i}}{\bar{x}_{2,i}}(x_{2,i})_T \right)^2 \frac{1}{\widetilde{x}_1(x_1)_T} \right) \frac{1}{(x_1)_T} \right| \\
\leq & K_2 \left| \left(\sum_{i=1}^{k} \left(\frac{\widetilde{x}_{2,i}^2}{\widetilde{x}_1} \right) \right) \right| \left| \max_i \left(\frac{1}{\bar{x}_{2,i}}(x_{2,i})_T \frac{1}{(x_1)_T} \right)^2 \right| \\
= & K_2 \left| \frac{\|(\widetilde{x_2})\|^2}{\widetilde{x}_1} \right| \left| \max_i \left(\frac{1}{\bar{x}_{2,i}}(x_{2,i})_T \frac{1}{(x_1)_T} \right)^2 \right| \\
\leq & K_2 \left| \frac{\|(\widetilde{x_2})\|^2}{\widetilde{x}_1} \right|
\end{aligned}$$

provided

$$\left(\frac{1}{\bar{x}_{2,i}}(x_{2,i})_T \frac{1}{(x_1)_T} \right) \leq 1 \text{ for } \forall i$$

Similarly

$$\left| O_{3,2,i}\left(\left\| \frac{\widetilde{x}_2}{\bar{x}_2}(x_2)_T \right\|^2 \frac{1}{\widetilde{x}_1(x_1)_T} \right) \frac{\bar{x}_{2,i}}{(x_{2,i})_T} \right|$$

equals to

$$\begin{aligned}
= & \left| O_{3,2,i}\left(\sum_{j=1}^{k} \left(\frac{\widetilde{x}_{2,j}}{\bar{x}_{2,j}}(x_{2,j})_T \right)^2 \frac{1}{\widetilde{x}_1(x_1)_T} \right) \frac{\bar{x}_{2,i}}{(x_{2,i})_T} \right| \\
\leq & K_3 \left| \left(\sum_{j=1}^{k} \left(\frac{\widetilde{x}_{2,j}}{\bar{x}_{2,j}}(x_{2,j})_T \right)^2 \frac{1}{\widetilde{x}_1(x_1)_T} \right) \frac{\bar{x}_{2,i}}{(x_{2,i})_T} \right| \\
\leq & K_3 \left| \left(\sum_{j=1}^{k} \left(\frac{\widetilde{x}_{2,j}^2}{\widetilde{x}_1} \right) \right) \right| \left| \max_j \left(\frac{(x_{2,j})_T^2}{\bar{x}_{2,j}^2} \frac{\bar{x}_{2,i}}{(x_{2,i})_T} \frac{1}{(x_1)_T} \right) \right| \\
= & K_3 \left| \frac{\|(\widetilde{x_2})\|^2}{\widetilde{x}_1} \right| \left| \max_j \left(\frac{(x_{2,j})_T^2}{\bar{x}_{2,j}^2} \frac{\bar{x}_{2,i}}{(x_{2,i})_T} \frac{1}{(x_1)_T} \right) \right| \\
\leq & K_3 \left| \frac{\|(\widetilde{x_2})\|^2}{\widetilde{x}_1} \right|
\end{aligned}$$

provided

$$\left(\frac{(x_{2,j})_T^2}{\bar{x}_{2,j}^2}\frac{\bar{x}_{2,i}}{(x_{2,i})_T}\frac{1}{(x_1)_T}\right) \leq 1 \text{ for } \forall i,j \tag{3.3.1.31}$$

The last condition, if true, will imply

$$\left(\frac{1}{\bar{x}_{2,i}}(x_{2,i})_T\frac{1}{(x_1)_T}\right) \leq 1 \text{ for } \forall i \tag{3.3.1.32}$$

from above. Finally, $O_1(|x_1| + \|x_2\|) = O_1\left(|\widetilde{x}_1(x_1)_T| + \|\widetilde{x}_2\frac{(x_2)_T}{\bar{x}_2}\|\right)$, and so is also Lipschitz continuous at 0 with respect to $|\widetilde{x}_1| + \|\widetilde{x}_2\|$ with at most same Lipschitz constant K_1 (since $x_1(t), \|x_2(t)\|$ are decreasing), and, in fact, the new Lipschitz constant will decrease exponentially compared to "original" K_1 (where the degree of exponent will correspond to the slowest decay rate in 3.3.1.24, 3.3.1.26, 3.3.1.28).

So if we can guarantee

$$\left(\frac{(x_{2,j})_T^2}{\bar{x}_{2,j}^2}\frac{\bar{x}_{2,i}}{(x_{2,i})_T}\frac{1}{(x_1)_T}\right) \leq 1 \text{ for } \forall i,j \leq 1$$

we can rewrite 3.3.1.30 as

$$\begin{cases} \dot{\widetilde{x}}_1 &= -\widetilde{x}_1 + (\eta + O_1(|\widetilde{x}_1| + \|\widetilde{x}_2\|))O_{2,1}\left(\frac{\|\widetilde{x}_2\|^2}{\widetilde{x}_1}\right) \\ &+ O_{3,1}\left(\frac{\|\widetilde{x}_2\|^2}{\widetilde{x}_1}\right) \\ \dot{\widetilde{x}}_{2,i} &= -\widetilde{x}_{2,i} + (\eta + O_1(|\widetilde{x}_1| + \|\widetilde{x}_2\|))(-\widetilde{x}_{2,i}) \\ &+ (\eta + O_1(|\widetilde{x}_1| + \|\widetilde{x}_2\|))O_{2,2,i}\left(\frac{\|\widetilde{x}_2\|^2}{\widetilde{x}_1}\right) + O_{3,2,i}\left(\frac{\|\widetilde{x}_2\|^2}{\widetilde{x}_1}\right), \forall i \end{cases} \tag{3.3.1.33}$$

with all of the $O_1, O_{i,j}, O_{i,j,k}$ again being Lipschitz continuous at 0 with constants $\widetilde{K_i} \leq K_i$, and moreover $\widetilde{K_1} < K_1$ decaying exponentially fast over $[0,T]$). The initial conditions 3.3.1.21 are also "the same" for system 3.3.1.33; now $((\widetilde{x}_1)_T = 1, \|(\widetilde{x}_2)_T\| \leq \|\bar{x}_2\|)$.

The re-scaling condition 3.3.1.32 (or 3.3.1.31) in particular gives us the wedge W_ϵ for all the solutions to 3.3.1.20, 3.3.1.21 with starting point $\bar{x}_2$ being below some threshold value (in Euclidean norm). It also tells us that any system 3.3.1.20

with initial point in W_ϵ can be re-scaled to produce 3.3.1.33, and thus we can apply to it the argument above. We can iterate the argument, letting $t \uparrow \infty$, provided 3.3.1.31 is satisfied.

To check that this condition can be met, observe

$$\left(\frac{(x_{2,j})_T^2}{\bar{x}_{2,j}^2}\frac{\bar{x}_{2,i}}{(x_{2,i})_T}\frac{1}{(x_1)_T}\right) \leq \frac{\bar{x}_{2,j}^2 e^{2(-1-\eta+\gamma)T}}{\bar{x}_{2,j}^2}\frac{\bar{x}_{2,i}}{\bar{x}_{2,i}e^{(-1-\eta-\gamma)T}}\frac{1}{1e^{(-1-\psi)T}} = e^{(-\eta+3\gamma+\psi)T}$$

and is ≤ 1 if we require that

$$\begin{aligned}(-\eta+3\gamma+\psi) &= -\eta+3((K_1\beta+K_1\|\bar{x}_2\|)\\ &+ (\eta+\beta K_1+K_1\|\bar{x}_2\|)K_2\alpha\left(\textstyle\sum_i|\bar{x}_{2,i}|\right)\\ &+ K_3\alpha\left(\textstyle\sum_i|\bar{x}_{2,i}|\right)) \qquad (3.3.1.34)\\ &+ ((\eta+K_1\beta+K_1\|\bar{x}_2\|)\alpha K_2+\alpha K_3)k\bar{x}_{2,j}^2\\ &< 0\end{aligned}$$

which, again, can be satisfied with a proper (sufficiently small) choice of $\|\bar{x}_2\|$ (i.e., ϵ). (We need to guarantee $-\eta+3K_1\beta<0$ and this is true under the assumptions of the lemma; compare with the assumption in Lemma 3.3.6, $K_1<\eta/2$).

Finally, let us examine how good are the established bounds in 3.3.1.24, 3.3.1.26, 3.3.1.28 as $t\uparrow\infty$. On $[0,T]$ we have

$$\|\bar{x}_2\|e^{(-1-\eta-\gamma)t} \leq \|x_2(t)\| \leq \|\bar{x}_2\|e^{(-1-\eta+\gamma)t}$$

$$\bar{x}_1 e^{(-1-\psi)t} \leq x_1(t) \leq \bar{x}_1 e^{(-1+\psi)t}$$

If we choose $\bar{x}_2$ to have strict inequality $(-1-\eta+\gamma)<(-1-\psi)$ which, again, is possible, then the ratio $\frac{\|x_2(t)\|}{x_1(t)}$ will also decay exponentially over $[0,T]$. Thus by iterating this bounds, we will arrive at progressively smaller $\|(\widetilde{x}_2)_T\|$ compared to $(\widetilde{x}_1)_T=1$. Recalling that $\widetilde{K_1}$ also decays exponentially on each interval of length T, we obtain our asymptotic result

$$x_1(t) \sim C_1 e^{-t}$$

$$\|x_2(t)\| \sim C_2 e^{-(1+\eta)t}$$

□

Bringing pieces together – asymptotic convergence for $e(t), t \uparrow \infty$

So far we have

$$(a) \quad \|N_{g_{\widetilde{e}}}(0) - y(\widetilde{e})\| \leq \tfrac{12\widehat{\gamma}}{t}\|NS_{g_{\widetilde{e}}}(0)\|^2 = \tfrac{12(n-m)\widehat{\gamma}}{\|e_\parallel\|}\|NS_{g_{\widetilde{e}}}(0)\|^2$$
$$(b) \quad N_{g_{\widetilde{e}}}(0) \equiv NS_{g_{\widetilde{e}}}(0) = \alpha(\widetilde{e})\left(-\widetilde{e}_\perp - \tfrac{(\widetilde{e}_\perp)^2}{\widetilde{e}_\parallel}\right)$$

where $\widetilde{e} = te$, $e \in \omega_{e_0,\epsilon}$ (for properly chosen $e_0 \in \mathbb{L}_{M^c}, \epsilon > 0$), $0 < t \leq 1$, $\widetilde{e} = \widetilde{e}_\perp + \widetilde{e}_\parallel$ (with $\widetilde{e}_\parallel = \tau\mathbf{1}$ for some $\tau > 0$) and

$$\left|\alpha(\widetilde{e}) - \frac{1}{m}\right| \leq K(\|\widetilde{e}_\parallel\| + \|\widetilde{e}_\perp\|)$$

Decomposing $e = e_\perp + e_\parallel$ we can write the ODE 3.2.1.1 for $e(t)$ as follows

$$\begin{aligned}
\dot{e} &= -e + \alpha(e)\left(-e_\perp - \tfrac{e_\perp^2}{e_\parallel}\right) + V_1 \\
&= -e + \left(\tfrac{1}{m} + O_1(\|e_\parallel\| + \|e_\perp\|)\right)(-e_\perp) \\
&\quad + \left(\tfrac{1}{m} + O_1(\|e_\parallel\| + \|e_\perp\|)\right)\left(\tfrac{-e_\perp^2}{e_\parallel}\right) + V_2 \\
&= -e + \left(\tfrac{1}{m} + O_1(\|e_\parallel\| + \|e_\perp\|)\right)(-e_\perp) \\
&\quad + \left(\tfrac{1}{m} + O_1(\|e_\parallel\| + \|e_\perp\|)\right)V_3 + V_4
\end{aligned}$$

(with $|O(y)| \leq |y|$,

$$\|V_1\|_\infty = O\left(\frac{12(n-m)\widehat{\gamma}}{\|e_\parallel\|}\left\|\alpha(e)\left(-e_\perp - \frac{e_\perp^2}{e_\parallel}\right)\right\|^2\right)$$

$$\|V_2\|_\infty = O\left(\frac{12(n-m)\widehat{\gamma}}{\|e_\parallel\|}(\alpha(e))^2\left(\|e_\perp\|^2 + \left\|\frac{e_\perp^2}{e_\parallel}\right\|^2 + 2\|e_\perp\|\left\|\frac{e_\perp^2}{e_\parallel}\right\|\right)\right)$$

$$\|V_3\|_\infty = O_2\left(\frac{\|e_\perp\|^2}{\|e_\parallel\|}\right)$$

$$\|V_4\|_\infty = O_3\left(\frac{\|e_\perp\|^2}{\|e_\parallel\|}\right)$$

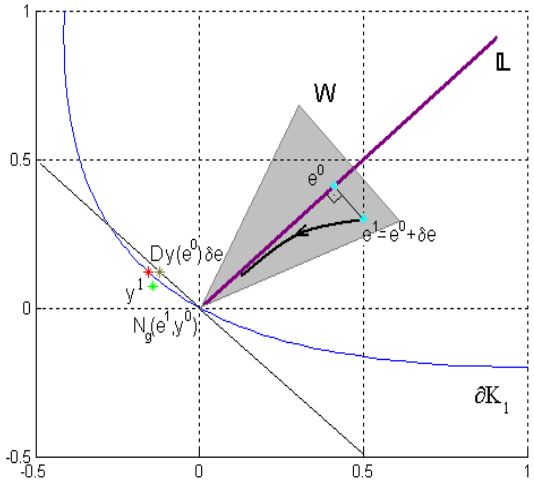

Figure 3.6: Shrink-Wrapping algorithm for LP, asymptotic behavior for $e(t)$

and $\|\cdot\|$ – Euclidian norm) where the second equality follows from triangle inequality and the third one follows from $e \in \mathbb{R}^{n-m}_{++}$, so that

$$\frac{e_\perp}{e_\parallel} \leq \mathbf{1}$$

and hence

$$\left\| \frac{e_\perp^2}{e_\parallel} \right\|^2 = \sum_i \left(\frac{1}{e_\parallel} \right)_i^2 {e_\perp}_i^4 \leq \sum_i {e_\perp}_i^2 = \|e_\perp\|^2$$

and from equivalence of norms in $\mathbb{R}^k$ (namely $\|\cdot\|_{l_1}$ and $\|\cdot\|_{l_2}$)

$$\begin{aligned} \|e_\perp\| \left\| \frac{e_\perp^2}{e_\parallel} \right\| &\leq M_1 \|e_\perp\| \left(\sum_i \left(\frac{1}{e_\parallel} \right)_i {e_\perp}_i^2 \right) \\ &\leq M_1 \|e_\perp\| \left(\sum_i |e_{\perp i}| \right) \leq M_2 \|e_\perp\| \|e_\perp\| = M_2 \|e_\perp\|^2 \end{aligned}$$

and finally with $\alpha(e)$ being finite and bounded in a properly chosen wedge $W_{e_0,\epsilon} = \{e : e \in t\omega_{e_0,\epsilon}, 0 \leq t \leq 1\}$.

Note that now the ODE determining the dynamics of $e(t)$ is in the form we want (i.e., compliant with the assumptions of Lemma 3.3.11). The Lipschitz constant

for $O_1(x)$ can be set arbitrarily small by picking a small neighborhood of $e = 0$ (wedge $W_{e_0,\epsilon}$). Re-scaling e back by introducing

$$\widetilde{x}_1 := \frac{e_{\|}}{\|(e_{\|})_0\|}$$

so that $\|(\widetilde{x}_1)_0\| \equiv 1$ and

$$\widetilde{x}_2 := \frac{e_{\perp}}{\|(e_{\|})_0\|}$$

(possibly $\ll 1$), we can choose a proper wedge $W_{e_0,\epsilon}$ for our result to hold by making an appropriate coordinate system rotation for $\widetilde{x}$ (i.e., align $\widetilde{x}_1$ with the first coordinate axes x_1, etc.) and applying Lemma 3.3.11. Thus the conclusion of Theorem 3.3.1 follows.

For an illustration based on Example 3.2.4, see Figure 3.6 (same setting as in Figure 3.5 with the wedge added).

3.3.2 Non-linear change of coordinates

First example and some ODE background

Recall Example 3.2.4: we considered the LP

$$\{\min_x [0,1,1]x : [1,1,1]x = 3, x \in \mathbb{R}^3_+\}$$

with optimal solution $x^* = (3,0,0)$ and introduced $y := (x_2, x_3)$ as a parametrization for the affine-feasible region; similarly for the affine-feasible hyperbolicity direction d we introduced $e := (d_2, d_3)$. We consider the quadratic relaxation of the LP

$$\{\min_x [0,1,1]x : [1,1,1]x = 3, x \in K_{2,d}\}$$

The optimality conditions for y (besides being a point on the boundary of $K_{\widehat{p}}$) are

$$\nabla_y \widehat{p} = \tau \mathbf{1}$$

with $\tau > 0$, that is,

$$z + Qy = \tau \mathbf{1}$$

where Q and z are defined as before.

Substituting for z

$$\begin{aligned} z &= 2 \begin{pmatrix} \widetilde{b} \\ 0 \end{pmatrix}^T [\mathbf{1}/d][\mathbf{1}\mathbf{1}^T - I][\mathbf{1}/d] \begin{pmatrix} \widetilde{A} \\ I_{2\times 2} \end{pmatrix} \\ &= 2 \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}^T [\mathbf{1}/d][\mathbf{1}\mathbf{1}^T - I][\mathbf{1}/d] \begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \frac{2}{d_1 d_2 d_3} \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}^T \begin{pmatrix} e_2 & e_1 \\ -e_2 & e_1 - e_2 \\ e_1 - e_1 & -e_1 \end{pmatrix} = \frac{6}{(3 - e_1 - e_2)e_1 e_2} \begin{pmatrix} e_2 \\ e_1 \end{pmatrix} \end{aligned}$$

the condition above (with respect to y) becomes

$$\begin{aligned} \tau \begin{pmatrix} 1 \\ 1 \end{pmatrix} &= \frac{6}{(3 - e_1 - e_2)e_1 e_2} \begin{pmatrix} e_2 \\ e_1 \end{pmatrix} \\ &+ \frac{2}{(3 - e_1 - e_2)e_1 e_2} \begin{pmatrix} -2e_2 & 3 - 2(e_1 + e_2) \\ 3 - 2(e_1 + e_2) & -2e_1 \end{pmatrix} y \end{aligned}$$

for some $\tau > 0$. Furthermore, for small e_1, e_2, we have seen that the $\det(Q) < 0$,

that is, Q is invertible and thus we can write (for e close to the origin)

$$\begin{aligned}
\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} &= \begin{pmatrix} -2e_2 & 3-2(e_1+e_2) \\ 3-2(e_1+e_2) & -2e_1 \end{pmatrix}^{-1} \left(\tau \begin{pmatrix} 1 \\ 1 \end{pmatrix} - 3 \begin{pmatrix} e_2 \\ e_1 \end{pmatrix} \right) \\
&= \frac{1}{\det(Q)} \begin{pmatrix} -2e_1 & 2(e_1+e_2)-3 \\ 2(e_1+e_2)-3 & -2e_2 \end{pmatrix} \left(\tau \begin{pmatrix} 1 \\ 1 \end{pmatrix} - 3 \begin{pmatrix} e_2 \\ e_1 \end{pmatrix} \right) \\
&= \frac{1}{\det(Q)} \left(\tau \begin{pmatrix} 2e_2-3 \\ 2e_1-3 \end{pmatrix} - 3 \begin{pmatrix} e_1(2e_1-3) \\ e_2(2e_2-3) \end{pmatrix} \right)
\end{aligned}$$

with

$$\det(Q) = -9 + 12(e_1+e_2) - 4(e_1^2+e_2^2) - 4e_1e_2$$

as before. We use boundary conditions, $\widehat{p}(x) = 0$, to determine τ. Let us introduce

$$\begin{aligned}
\omega &:= det(Q) = -9 + 12(e_1+e_2) - 4(e_1^2+e_2^2) - 4e_1e_2, \\
V &:= \begin{pmatrix} 2e_2-3 \\ 2e_1-3 \end{pmatrix}, \quad U := -3 \begin{pmatrix} e_1(2e_1-3) \\ e_2(2e_2-3) \end{pmatrix},
\end{aligned}$$

Then we can write $[y_1; y_2] = \frac{1}{\omega}(\tau V + U)$. The boundary condition becomes

$$\begin{aligned}
0 &= 6\begin{pmatrix} e_2 \\ e_1 \end{pmatrix}^T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}^T \begin{pmatrix} -2e_2 & 3-2(e_1+e_2) \\ 3-2(e_1+e_2) & -2e_1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\
&= 6\begin{pmatrix} e_2 \\ e_1 \end{pmatrix}^T \left(\frac{1}{\omega}(\tau V + U)\right) \\
&+ \left(\frac{1}{\omega}(\tau V + U)\right)^T \begin{pmatrix} -2e_2 & 3-2(e_1+e_2) \\ 3-2(e_1+e_2) & -2e_1 \end{pmatrix} \left(\frac{1}{\omega}(\tau V + U)\right) \\
&= \frac{1}{\omega^2}\tau^2 \left(V^T \begin{pmatrix} -2e_2 & 3-2(e_1+e_2) \\ 3-2(e_1+e_2) & -2e_1 \end{pmatrix} V \right) \\
&+ \frac{1}{\omega^2}\tau \left(6\omega \begin{pmatrix} e_2 \\ e_1 \end{pmatrix}^T V + 2V^T \begin{pmatrix} -2e_2 & 3-2(e_1+e_2) \\ 3-2(e_1+e_2) & -2e_1 \end{pmatrix} U \right) \\
&+ \frac{1}{\omega^2} \left(U^T \begin{pmatrix} -2e_2 & 3-2(e_1+e_2) \\ 3-2(e_1+e_2) & -2e_1 \end{pmatrix} U \right) \\
&= \frac{1}{\omega^2}\tau^2 V^T \mathbf{1}\omega + \frac{1}{\omega^2}\tau(6\omega(e_2(2e_2-3)+e_1(2e_1-3)) + 2U^T\mathbf{1}\omega) \\
&+ \frac{1}{\omega^2} U^T \begin{pmatrix} -3e_2 \\ -3e_1 \end{pmatrix} \omega \\
&= \frac{1}{\omega}\tau^2(2(e_1+e_2)-6) + \frac{9}{\omega}e_1e_2(2(e_1+e_2)-6) \\
&+ \frac{1}{\omega}\tau(6(e_2(2e_2-3)+e_1(2e_1-3)) - 6(e_2(2e_2-3)+e_1(2e_1-3))) \\
&= \frac{1}{\omega}(\tau^2 + 9e_1e_2)(2(e_1+e_2)-6)
\end{aligned}$$

Since we require $\tau > 0$, we pick

$$\tau = 3(e_1e_2)^{1/2}$$

Finally

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \frac{3}{-9+12(e_1+e_2)-4(e_1^2+e_2^2)-4e_1e_2} \left(\begin{pmatrix} (e_1e_2)^{1/2}(2e_2-3) \\ (e_1e_2)^{1/2}(2e_1-3) \end{pmatrix} - \begin{pmatrix} e_1(2e_1-3) \\ e_2(2e_2-3) \end{pmatrix} \right)$$

We will use this to understand local behavior of $e(t)$ when e gets "close" to the optimal LP solution $e = 0$ (i.e., x^*). Recall the ODE 3.2.1.1 defining the dynamics for $e(t), t \geq 0$, is given by

$$\dot{e}(t) = y(e(t)) - e(t)$$
$$e(0) = e$$

At this point we give a few definitions and the result to be used that we borrow from dynamical systems (see, for example, [18]). Consider a differential equation

$$\dot{x} = f(x), \quad f : W \to \mathbb{R}^n \text{ is in } C^1, \quad W \subset \mathbb{R}^n \text{ open}$$

Definition 3.3.12. A point $\bar{x} \in W$ is called an *equilibrium (stationary) point* if $f(\bar{x}) = 0$.

Clearly, the constant function $x(t) \equiv \bar{x}$ is a solution to this ODE.

Suppose 0 is such an equilibrium point. Think of the derivative $Df(0)$ of f at 0 as a linear vector field which approximates f near 0. If all the eigenvalues of $Df(0)$ have nonnegative real parts, we call 0 a sink. More generally:

Definition 3.3.13. An equilibrium $\bar{x}$ is a *sink* if all eigenvalues of $Df(\bar{x})$ have nonnegative real parts.

In words, one can use just the linear part of the equation around the sink to understand how it acts locally (in a neighborhood of the sink). We denote the solution to this system as $\phi_t(x)$ (meaning a function depending on time parameter t such that $\dot{\phi}_t(x) = f(\phi_t(x))$ and corresponding to the initial condition $\phi_0(x) = x$).

Theorem 3.3.14. *Let $\bar{x}$ be a sink. Suppose every eigenvalue of $Df(\bar{x})$ has real part less than $-c$, $c > 0$. Then there is a neighborhood $U \subset W$ of $\bar{x}$ such that*

(a) $\phi_t(x)$ is defined and in U for all $x \in U$, $t > 0$

(b) There is an Euclidean norm on $\mathbb{R}^n$ *such that*

$$\|\phi_t(x) - \bar{x}\| \leq e^{-tc}\|x - \bar{x}\|$$

for all $x \in U$, $t \geq 0$

(c) For any norm on $\mathbb{R}^n$, *there is a constant* $B > 0$ *such that*

$$\|\phi_t(x) - \bar{x}\| \leq Be^{-tc}\|x - \bar{x}\|$$

for all $x \in U$, $t \geq 0$

In particular, $\phi_t(x) \to \bar{x}$ *as* $t \to \infty$ *for all* $x \in U$.

Proof. See [18], Chapter 9. □

What does this have to do with our problem? Seemingly, the point $e = 0$ is "almost" a stationary point for the ODE

$$\dot{e} = y(e) - e$$

The Jacobian of $y(e)$ for $e \in \mathbb{L}_{M^c}$ suggests that if $e = 0$ was a sink (i.e., was in the interior of the domain of $e \mapsto y(e) - e$), then the result above would be applicable (recall the evaluation of the Jacobian of $y(e)$ on $\mathbb{L}_{M^c}$) and we would have our asymptotic convergence result (namely, the ODE 3.2.1.1 would locally act as a its linear counterpart). Therefore, we would like to extend $y(e)$ beyond positive orthant (its current domain) in a C^1 fashion (so that 0 will be in its open domain). Currently the domain for $y(e)$ is $\mathbb{R}^{(n-m)}_{++}$.

Observe that even in our simple example we cannot make a straightforward C^1 extension of this function beyond the positive orthant. The reason for this is having a square root in the enumerator, that is, having a term

$\sqrt{e_1 e_2}$: this term will result in a singularity of the first derivative when, for example, e_2 stays fixed and $e_1 \downarrow 0$. Also, note that if both e_1 and e_2 go to zero along some line (in $\mathbb{R}^{(n-m)}_{++}$) this will not cause a problem anymore. This suggests we try switching to the polar coordinates and extending $y(e)$ "through" 0.

Introduce polar coordinates in the plane for $y, e \in \mathbb{R}^2_{++}$:

$$\begin{cases} y_1 = \rho \cos\phi \\ y_2 = \rho \sin\phi \end{cases}$$

and

$$\begin{cases} e_1 = \rho \cos\phi \\ e_2 = \rho \sin\phi \end{cases}$$

for $\rho \in \mathbb{R}_{++}$ and $\phi \in (0, \pi/2)$. In this new coordinates we can write

$$\det(Q) = -9+12(e_1+e_2)-4(e_1^2+e_2^2)-4e_1e_2 = -9+12\rho(\cos\phi+\sin\phi)-4\rho^2-2\rho\sin 2\phi$$

and letting $\rho > 0$ (for now) since

$$\begin{aligned} \sqrt{e_1 e_2}(2e_2 - 3) - e_1(2e_1 - 3) &= \rho\sqrt{\sin\phi\cos\phi}(2\rho\sin\phi - 3) - \rho\cos\phi(2\rho\cos\phi - 3) \\ &= 3\rho\left(\cos\phi - \sqrt{\frac{\sin 2\phi}{2}}\right) \\ &+ 2\rho^2\left(\sqrt{\frac{\sin 2\phi}{2}}\sin\phi - \cos^2\phi\right) \end{aligned}$$

(and similarly for the second component of $y(e)$) we have

$$\begin{aligned} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} &= \frac{3}{\det(Q)}\left(\begin{pmatrix} (e_1e_2)^{1/2}(2e_2-3) \\ (e_1e_2)^{1/2}(2e_1-3) \end{pmatrix} - \begin{pmatrix} e_1(2e_1-3) \\ e_2(2e_2-3) \end{pmatrix}\right) \\ &= \frac{3}{-9+12\rho(\cos\phi+\sin\phi)-4\rho^2-2\rho\sin 2\phi} \\ &\quad \begin{pmatrix} 3\rho\left(\cos\phi - \sqrt{\frac{\sin 2\phi}{2}}\right) + 2\rho^2\left(\sqrt{\frac{\sin 2\phi}{2}}\sin\phi - \cos^2\phi\right) \\ 3\rho\left(\sin\phi - \sqrt{\frac{\sin 2\phi}{2}}\right) + 2\rho^2\left(\sqrt{\frac{\sin 2\phi}{2}}\cos\phi - \sin^2\phi\right) \end{pmatrix} \end{aligned}$$

Note that y, as a function of ρ and ϕ, is well-defined for $-\epsilon < \rho < \epsilon$ (with $\epsilon > 0$ small) and $\phi \in (0, \pi/2)$. Moreover, the expression above is obviously a C^1 function on this open domain containing the point $(\rho, \phi) = (0, \pi/4)$ at which it vanishes.

Apply the same coordinate change to our ODE 3.2.1.1. Consider

$$\begin{cases} \dot{e}_1 = \dot{\rho}\cos\phi - \rho\dot{\phi}\sin\phi \\ \dot{e}_2 = \dot{\rho}\sin\phi + \rho\dot{\phi}\cos\phi \end{cases}$$

so

$$\begin{cases} \dot{\rho}\cos\phi - \rho\dot{\phi}\sin\phi = -\rho\cos\phi + y_1 \\ \dot{\rho}\sin\phi + \rho\dot{\phi}\cos\phi = -\rho + y_2 \end{cases}$$

Rearranging terms (say, multiply first and second equation by $\cos\phi$ and $\sin\phi$ respectively and add them up to get $\dot{\rho}$; substitute for $\dot{\rho}$ to get $\dot{\phi}$) we get

$$\begin{cases} \dot{\rho} = -\rho + y_1\cos\phi + y_2\sin\phi \\ \dot{\phi} = \frac{1}{\rho}(y_2\cos\phi - y_1\sin\phi) \end{cases}$$

(note that division by ρ does not cause a problem as $\rho \to 0$, since y, as a function of ρ, has ρ in the enumerator, see above).

Finally, we can write

$$\begin{cases} \dot{\rho} &= -\rho + \frac{3}{\det(Q)}\cos\phi\left(3\rho\left(\cos\phi - \sqrt{\frac{\sin 2\phi}{2}}\right) + 2\rho^2\left(\sqrt{\frac{\sin 2\phi}{2}}\sin\phi - \cos^2\phi\right)\right) \\ &\quad + \frac{3}{\det(Q)}\sin\phi\left(3\rho\left(\sin\phi - \sqrt{\frac{\sin 2\phi}{2}}\right) + 2\rho^2\left(\sqrt{\frac{\sin 2\phi}{2}}\cos\phi - \sin^2\phi\right)\right) \\ \dot{\phi} &= \frac{3}{\det(Q)}\cos\phi\left(3\left(\sin\phi - \sqrt{\frac{\sin 2\phi}{2}}\right) + 2\rho\left(\sqrt{\frac{\sin 2\phi}{2}}\cos\phi - \sin^2\phi\right)\right) \\ &\quad - \frac{3}{\det(Q}\sin\phi\left(3\left(\cos\phi - \sqrt{\frac{\sin 2\phi}{2}}\right) + 2\rho\left(\sqrt{\frac{\sin 2\phi}{2}}\sin\phi - \cos^2\phi\right)\right) \end{cases}$$

with

$$\det(Q) = -9 + 12\rho(\cos\phi + \sin\phi) - 4\rho^2 - 2\rho\sin 2\phi$$

The point $(\rho, \phi) = (0, \pi/4)$ is a stationary point for our (new) system. Moreover, the domain on which the function on the right-hand side of the equation is C^1 can be taken as $-\epsilon < \rho < \epsilon$ (for some small $\epsilon > 0$) and $\phi \in (0, \pi/2)$.

Note that $\rho = 0$ and $\phi \in (0, \pi/2)$ is a unique stationary point (the condition for $\dot{\phi} = 0$ translates into $0 = \cos\phi\sqrt{(\sin 2\phi)/2} - \sin\phi\sqrt{(\sin 2\phi)/2}$, that is, $\sin\phi = \cos\phi$). This is somewhat counterintuitive, since for $\rho = 0$ in the original problem $e = 0$, which is the optimal solution for (P), regardless of the polar angle ϕ.

Also, observe that this once again illustrates the existence of the central line in this particular setting: for small ρ and $\phi = \pi/4$ (corresponding to the diagonal of $\mathbb{R}^2_+$), $\dot{\phi} = 0$, that is, if we were on the central line to begin with, we will necessarily stay there until we reach the origin.

Evaluating the derivative of the right-hand side of this differential equation at $(\rho, \phi) = (0, \pi/4)$ gives us $-I$. Therefore, $(0, \pi/4)$ is a sink and the local dynamics of the system can be analyzed using the theorem quoted above. In particular, as an immediate consequence, we observe that if the initial point e is "close enough" to the central line and "close enough" to the origin (both $\rho > 0$ and $|\phi - \pi/4|$ are small), then we will be pulled into 0 exponentially fast in polar coordinates, thus asymptotically to $\mathbb{L}_{M^c}$ in Euclidean coordinates (the polar angle will tend to $\pi/4$ exponentially fast with $t \uparrow \infty$). The final observation also suggests to us the shape of the neighborhood of $\mathbb{L}_{M^c}$ where these convergence properties are exhibited. See Figure 3.7. Here $\mathrm{Dom}(f)$ and $\mathrm{Dom}(\widetilde{f})$ correspond to domains of the original function $y(e)$ (or $(y(e) - e)$) in Euclidean and polar coordinates respectively; U is a neighborhood of exponential convergence for the polar radius and the polar angle (say, a ball in l_1-norm) and its equivalent in Euclidean coordinate system (a "two-sided wedge").

Establishing a sink: from $\mathbb{R}^n$ to polar coordinates and back

Can we apply the same line of reasoning to general $(P), (P(d))$?

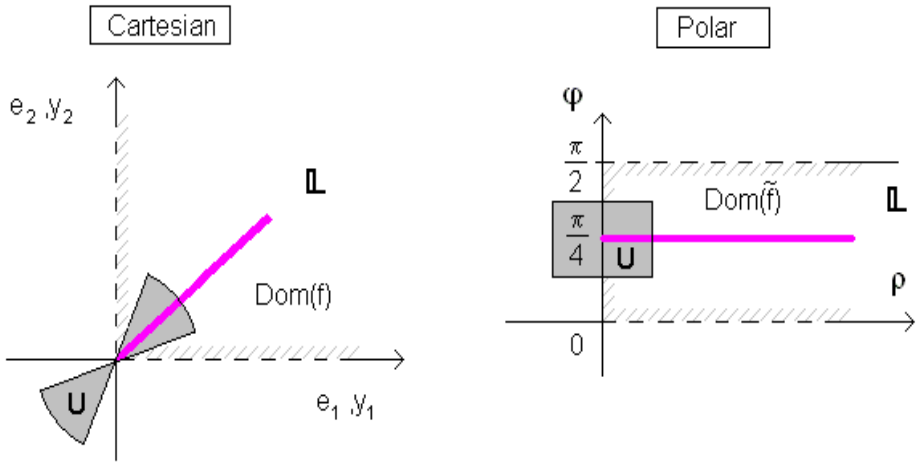

Figure 3.7: Shrink-Wrapping algorithm for LP, cartesian and polar domains for $y(e)$

Similarly to the case of $\mathbb{R}^2$, we introduce the change of coordinates (from spherical to Euclidean) for $\mathbb{R}^n$ as follows:

$$\begin{cases} x_1 = \rho \prod_{i=1}^{n-1} \cos\phi_i \\ x_i = \rho \sin\phi_{i-1} \prod_{j=i}^{n-1} \cos\phi_j, \text{ for } i \geq 2 \end{cases}$$

that is, for example, in $\mathbb{R}^3$

$$\begin{cases} x_1 = (\rho\cos\phi_1)\cos\phi_2 \\ x_2 = (\rho\sin\phi_1)\cos\phi_2 \\ x_3 = (\rho)\sin\phi_2 \end{cases}$$

Denote this map

$$\Psi_n : \mathbb{R}_+ \times [-\pi,\pi)^{n-1} \to \mathbb{R}^n$$

$$(\rho,\phi) \mapsto x$$

Lemma 3.3.15. *Let $\widetilde{f} : \mathbb{R}\times\mathbb{R}^{n-1} \to \mathbb{R}\times\mathbb{R}^{n-1}$ be a C^1 function in spherical coordinates (ρ,ϕ) with domain $W \subseteq \mathbb{R}\times\mathbb{R}^{n-1}$ such that $W \supset U = (-\epsilon,\epsilon)\times$*

$(-\gamma,\gamma)^{n-1}$ for some $\epsilon,\gamma>0$ ($\gamma<\pi$). Suppose $\nabla_{\rho,\phi}\Psi_n(\rho,\phi)^{-1}f(\Psi(\rho,\phi))=\widetilde{f}(\rho,\phi)$ on $V=(0,\epsilon)\times(-\gamma,\gamma)^{n-1}$ for some C^1 function $f:\Psi_n(V)\to\mathbb{R}^n$. Suppose that $\forall x\in(0,\epsilon)\times\{0\}^{n-1}$ the Jacobian of f, $\nabla_x f(x)$, has diagonal structure with finite limit as $x=(x_1,0,0,\ldots,0)\to\mathbf{0}$ and also $f(x)\to\mathbf{0}$. Then the Jacobian of $\widetilde{f}$ satisfies

$$\nabla_{\rho,\phi}\widetilde{f}(0,\mathbf{0})=\begin{pmatrix}\frac{\partial f_1}{\partial x_1} & & & & \\ & (\frac{\partial f_2}{\partial x_2}-\frac{\partial f_1}{\partial x_1}) & & & \\ & & (\frac{\partial f_3}{\partial x_3}-\frac{\partial f_1}{\partial x_1}) & & \\ & & & \ddots & \\ & & & & (\frac{\partial f_{n+1}}{\partial x_{n+1}}-\frac{\partial f_1}{\partial x_1})\end{pmatrix}$$

and

$$\widetilde{f}(0,\mathbf{0})=(0,\mathbf{0})$$

Proof. Mostly computational, see Appendix C. □

Theorem 3.3.16. *Consider ODE*

$$\dot{x}=f(x)$$

with $f:\Psi_n(V)\to\mathbb{R}^n$ as above. Assume $\forall x\in\Psi(V),\nabla_x f(x)\prec 0$ (i.e., $\frac{\partial f_i}{\partial x_i}<0,\forall i$) and $\left|\frac{\partial f_1}{\partial x_1}\right|\leq\left|\frac{\partial f_j}{\partial x_j}\right|-\nu,\forall j\neq 1$, for some $\nu>0$. Then the corresponding (equivalent) ODE in spherical coordinates

$$\begin{pmatrix}\dot{\rho}\\ \dot{\phi}\end{pmatrix}=\widetilde{f}\begin{pmatrix}\rho\\ \phi\end{pmatrix}$$

(with $\widetilde{f}$ as above) will have a sink at $(\rho,\phi)=(0,\mathbf{0})$.

Proof. By the chain rule

$$\dot{x}=\nabla_{\rho,\phi}\Psi_n(\rho,\phi)\begin{pmatrix}\dot{\rho}\\ \dot{\phi}\end{pmatrix}$$

where

$$\nabla_{\rho,\phi}\Psi_n(\rho,\phi) = \begin{pmatrix} \frac{\partial x_1}{\partial \rho} & \frac{\partial x_1}{\partial \phi_1} & \frac{\partial x_1}{\partial \phi_2} & \dots \\ \frac{\partial x_2}{\partial \rho} & \frac{\partial x_2}{\partial \phi_1} & \frac{\partial x_2}{\partial \phi_2} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

is the Jacobian of Ψ_n.

Now having $x = \Psi_n(\rho, \phi)$ we can rewrite ODE

$$\dot{x} = f(x)$$

as

$$\nabla_{\rho,\phi}\Psi_n(\rho,\phi) \begin{pmatrix} \dot{\rho} \\ \dot{\phi} \end{pmatrix} = f(\Psi_n(\rho,\phi))$$

and if the Jacobian of Ψ_n is invertible we can write

$$\begin{pmatrix} \dot{\rho} \\ \dot{\phi} \end{pmatrix} = \nabla_{\rho,\phi}\Psi_n(\rho,\phi)^{-1} f(\Psi_n(\rho,\phi)) = \widetilde{f}(\rho,\phi)$$

an equivalent ODE in spherical coordinates. We will refer to $\widetilde{f}$ as *the spherical coordinate analogue* of f.

Now from the lemma above it follows that the point $(\rho, \phi) = (0, \mathbf{0})$ is a stationary point for this new ODE and, moreover, $\nabla_{\rho,\phi}\widetilde{f}(0, \mathbf{0}) \prec 0$, so it is a sink. □

Regarding our particular setting we can state the following.

Corollary 3.3.17. *For the ODE 3.2.1.1, if there exists a C^1 extension of $\widetilde{f}$, the spherical coordinate analogue of $f(e) := (y(e) - e)$, beyond the positive orthant (for $(\rho, \phi) \in (\epsilon, \epsilon) \times (\pi/4 - \gamma, \pi/4 + \gamma)^{n-m-1}$ for some $\epsilon, \gamma > 0$), then the point $(\rho, \phi) = (0, \pi/\mathbf{4})$ is a sink for $y(e)$ in spherical coordinates. Moreover, $e(t)$ will converge asymptotically to $\mathbb{L}_{M^c}$ as $t \uparrow \infty$ in Euclidean coordinate system for any starting point in some (properly chosen) wedge $W_{e_0,\epsilon}$.*

Proof. In order to put us in the setting of the theorem above apply the rotation to $\mathbb{R}^{n-m}_{++}$, the domain of $y(e)$ (or $y(e)-e$), that will set x_1 axis collinear with the central line $\mathbb{L}_{M^c} \parallel \mathbf{1}$ (the rotation will correspond to the eigenvectors of the Jacobian of $y(e)$ on $\mathbb{L}_{M^c}$). Note that the eigenvalues of

$$\lim_{t\downarrow 0}\left(-\frac{E_m}{E_{m-1}}(x^*_M/d_M(t\mathbf{1}))\left(I-\frac{\mathbf{1}\mathbf{1}^T}{n-m}\right)\right)-I=-\frac{1}{m}\left(I-\frac{\mathbf{1}\mathbf{1}^T}{n-m}\right)-I$$

are

$$-1,-1\frac{1}{m},-1\frac{1}{m},\ldots,-1\frac{1}{m}$$

The theorem above implies exponential convergence for $(\rho(t),\phi(t))$ in the proper neighborhood of $(0,\mathbf{0})$ (that corresponds to $(0,\pi/\mathbf{4})$ in the original coordinates) and thus gives us asymptotic convergence in Euclidian coordinates (same as in Example 3.2.4 at the beginning of this section). □

Remark 3.3.18. To establish the conclusion of the corollary above we did not require the knowledge of $y(e)$, as opposed to what we did in Example 3.2.4. The only part that is missing to make this a complete argument is to show that such $\widetilde{f}\in C^1$ does exist (We conjecture that we need to extend $\widetilde{f}$ for $\rho\geq 0$ only and this indeed can be done).

The switch to spherical coordinates exhibits an interesting effect of "damping" the diagonal entries of the Jacobian (the linearization) $\nabla_{\rho,\phi}\widetilde{f}$ by the slowest decaying exponent $\frac{\partial f_1}{\partial x_1}$, which is also consistent with the observation one can draw from the cartesian setting (recall, for small $\phi, \sin\phi\approx\phi$, and if $\|e_\parallel\|\gg\|e_\perp\|$, then $(e_\perp)_i\approx\sin\phi_i e_\parallel$).

3.4 Discrete setting and the rate of convergence

Consider the free variables coordinate system as before. Take $e = e_{\|} + e_{\perp}$ with $e_{\|} \in \mathbb{L}_{M^c}$ and $e_{\perp}$ being its orthogonal complement. Recall that if $e \in \mathbb{L}_{M^c}$, then $y(e) = 0$ corresponding to x^*, the optimal solution for the LP. Given a current iterate e_i (with the initial iterate e_0 in a properly chosen wedge $W_{\widehat{e}_0,\epsilon}$) we will attempt to place the next iterate e_{i+1} onto the central line $\mathbb{L}_{M^c}$ as follows (that is, get rid of the $e_{\perp}$ component): if $y(e) \approx -\frac{1}{m}e_{\perp}$, then setting $e_{i+1} := \frac{my_i+e_i}{1+m}, i = 0, 1, 2, \ldots$, we would hope to converge to the central line $\mathbb{L}_{M^c}$ fairly quickly.

We show that following this naive scheme indeed gives us the superlinear convergent series $\{y_i\}_{i\geq 0}$ with a limit point being the optimal LP solution under certain assumption on the starting point e_0. See Figure 3.8, where one such iteration is illustrated based on Example 3.2.4. Given the initial iterate (e^0, y^0), depicted are the "true" next iterate e^1 based on actual y^0 and the "ideal" would-be iterate which belongs to the central line $\mathbb{L}_{M^c}$, based on the linearization part of y^0, $\delta y(e^0)$.

Remark 3.4.1. Recall that we are assuming that given e one can easily compute $y(e)$, the corresponding solution to the relaxed problem $(P([d_M(e); e]))$ (possibly, this assumption can be further lifted by implementing a Newton-like procedure for updating $y(e_i)$ iterates).

We choose the following coordinate system in $\mathbb{R}^{n-m}$: take $(n-m)$ unit vectors such that for any vector e its first coordinate e_1 will correspond to $\|e_{\|}\|$, and the remaining $(n-m)-1$ components will represent the remaining $e_{\perp}$ part of e. From now on we will refer to the first component of a vector e as $e_{\|} \in \mathbb{R}$ and to the remaining $(n-m)-1$ components of e as $e_{\perp} \in \mathbb{R}^{(n-m)-1}$ (we will not introduce

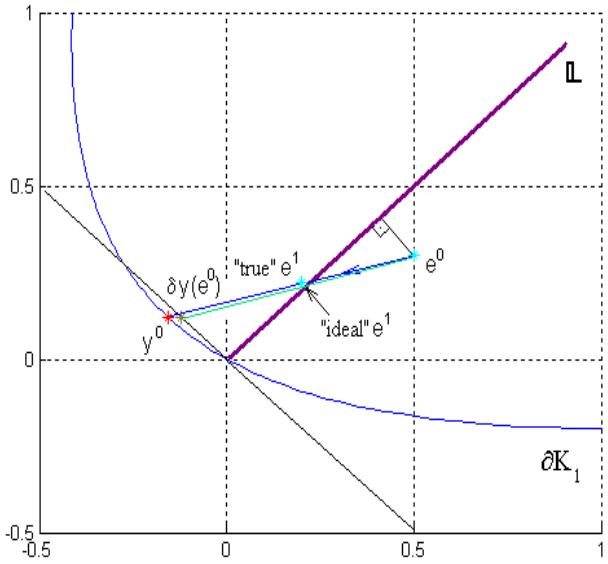

Figure 3.8: Shrink-Wrapping algorithm for LP, "fast" discrete convergence for $\{e_i\}$

new notation for e). For the i^{th} iterate e_i we will be writing $(e_{\|})_i$ and $(e_{\perp})_i$ for its first and last $(n-m-1)$ components.

Suppose the initial point e_0 is chosen such that the following is true

$$y(e) = \left(\frac{1}{m} + O_1\left(|e_{\|}| + \|e_{\perp}\|\right)\right)(-e_{\perp}) + V \tag{3.4.0.1}$$

where

$$\|V\|_{\infty} \leq O_2\left(\frac{\|e_{\perp}\|^2}{e_{\|}}\right)$$

and $e = e_{\|} + e_{\perp}$ with $e_{\|} = e_{\text{range}(c_Y^T)} = e_{\text{range}(\mathbf{1}^T)}$ as before ($e_{\|} = t\mathbf{1}, t > 0$), and O_1, O_2 Lipschitz continuous at 0. Such a choice is indeed possible, e.g., take $e \in W_{\widehat{e}_0,\epsilon}$ as in Theorem 3.3.1. If all the subsequent iterates stay in this wedge $W_{\widehat{e}_0,\epsilon}$, then 3.4.0.1 is true $\forall e_i$; the term corresponding to $\left(\frac{1}{m} + O_1(|e_{\|}| + \|e_{\perp}\|)\right)$ is bounded and thus we can combine two of the higher order terms into one $O_2\left(\frac{\|e_{\perp}\|^2}{e_{\|}}\right)$.

W.l.o.g. we may assume that 3.4.0.1 is true as long as

$$\frac{\|e_\perp\|}{e_\parallel} \leq \eta, e_\parallel \leq 1 \tag{3.4.0.2}$$

($\eta < 1$), $O_1(\cdot)$ and $O_2(\cdot)$ are both Lipschitz continuous at 0 with constants $\epsilon \ll 1$ and M respectively, and the initial point e_0 is chosen such that $(e_\parallel)_0 = 1$ (note that ϵ can be assumed arbitrarily small by taking a small enough wedge $W_{\widehat{e}_0,\epsilon}$ and re-scaling it so that $(e_\parallel)_0 = 1$).

Let us analyze what the next iterate e_1 would be, given by the recursion

$$e_{i+1} := \frac{my(e_i) + e_i}{1+m} \tag{3.4.0.3}$$

and what additional assumptions we need to impose to get the desired convergence result.

In particular, our goal is to prove that the initial point e_0 can be chosen such that the following estimates hold true $\forall i \geq 0$:

$$\left(\frac{1}{2(m+1)}\right)^i \leq (e_\parallel)_i \leq \left(\frac{2}{(m+1)}\right)^i \tag{3.4.0.4}$$

$$\left(\frac{\|e_\perp\|}{e_\parallel}\right)_i \leq \eta \left(\frac{2}{m+1}\right)^{1+2+\cdots+i} = \eta \left(\frac{2}{m+1}\right)^{\frac{i(i+1)}{2}} \tag{3.4.0.5}$$

$$\|(e_\perp)_i\| \leq \eta \left(\frac{2}{m+1}\right)^{1+2+\cdots+i+(i+1)-1} = \eta \left(\frac{2}{m+1}\right)^{\frac{(i+1)(i+2)}{2}-1} \tag{3.4.0.6}$$

giving us R-superlinear rate of convergence for $\{y_i\}_{i\geq 0}$ (the implications of this estimates for y_i will be derived at the very end of this discussion; for now one can think of $y_i \approx -\frac{1}{m}(0, (e_\perp)_i)$).

Note that 3.4.0.4, 3.4.0.5 imply 3.4.0.6, so we just need to demonstrate 3.4.0.4, 3.4.0.5. From 3.4.0.5 it follows that

$$\left(\frac{\|e_\perp\|}{e_\parallel}\right)_i \leq \eta$$

and thus 3.4.0.1 is applicable.

Observe

$$e_1 = \frac{1}{1+m} \begin{pmatrix} e_{\|} + mO_2\left(\frac{\|e_\perp\|^2}{e_{\|}}\right) \\ mO_1(|e_{\|}| + \|e_\perp\|)(-e_\perp) + mV \end{pmatrix}_0 \tag{3.4.0.7}$$

where

$$\|V\|_\infty \leq O_2\left(\frac{\|e_\perp\|^2}{e_{\|}}\right)_0$$

(to simplify the notation we put one subindex outside of brackets surrounding the whole expression to indicate the dependence on the components of e_0; here V is a vector of corresponding dimension to $e_\perp$, i.e., in $\mathbb{R}^{(n-m)-1}$).

One can easily choose a starting point e_0 such that

$$2 \geq \left(e_{\|} + mO_2\left(\frac{\|e_\perp\|^2}{e_{\|}}\right)\right)_0 \geq \frac{1}{2}$$

This will necessarily be true if

$$1 - mM\eta^2 \geq \frac{1}{2} \tag{3.4.0.8}$$

(recall $(e_{\|})_0 = 1$ and $\left(\frac{\|e_\perp\|}{e_{\|}}\right)_0 \leq \eta$) and this can be easily satisfied be choosing η sufficiently small, namely,

$$\eta \leq \sqrt{\frac{1}{2mM}}$$

With this choice of e_0 (i.e., η) from 3.4.0.7 we get

$$\begin{aligned} \left(\frac{\|e_\perp\|}{e_{\|}}\right)_1 &\leq 2\left\|\left(mO_1(|e_{\|}| + \|e_\perp\|)(-e_\perp) + mV\right)_0\right\| \\ &\leq 2\left\|(m\epsilon(1+\eta)(-e_\perp))_0\right\| + \|mM\eta^2\mathbf{1}\| \\ &\leq 2(2m\epsilon\eta + m\eta^2 M\sqrt{n-m}) \\ &\leq 2\eta(2m\epsilon + m\eta M\sqrt{n-m}) \leq \tfrac{2\eta}{1+m} \end{aligned} \tag{3.4.0.9}$$

provided

$$2m\epsilon + m\eta M\sqrt{n-m} \leq \frac{1}{1+m} \tag{3.4.0.10}$$

which, again, can be easily met for small enough $\eta, \epsilon > 0$ (recall that we can assume ϵ to be arbitrarily small).

If 3.4.0.8, 3.4.0.10 are both true we have

$$\left(\frac{1}{2(m+1)}\right) \leq (e_{\|})_1 \leq \left(\frac{2}{(m+1)}\right) \tag{3.4.0.11}$$

$$\left(\frac{\|e_{\perp}\|}{e_{\|}}\right)_1 \leq \eta \left(\frac{2}{m+1}\right) \tag{3.4.0.12}$$

$$\|(e_{\perp})_1\| \leq \eta \left(\frac{2}{m+1}\right)^2 \tag{3.4.0.13}$$

So assume we start with such ϵ, η. This forms the induction base for showing 3.4.0.4, 3.4.0.5, 3.4.0.6 (although these estimates are obviously true for $i = 0$, it is an illustrative exercise to show it is true for $i = 1$).

Assume 3.4.0.4, 3.4.0.5, hold for some $i = k \geq 0$, we want to show that this is true for $i = (k+1)$ as well. Introducing

$$\widetilde{e}_0 = (\widetilde{e}_{\perp}, \widetilde{e}_{\|})_0 := \frac{1}{(e_{\|})_k}(e_{\|}, e_{\perp})_k = \left(1, \left(\frac{e_{\perp}}{e_{\|}}\right)_k\right)$$

by simply re-scaling e_k we note that

$$\left\| \left(\frac{\widetilde{e}_{\perp}}{\widetilde{e}_{\|}}\right)_0 \right\| = \|(\widetilde{e}_{\perp})_0\| = \eta \left(\frac{2}{m+1}\right)^{\frac{k(k+1)}{2}} \leq \eta \left(\frac{2}{m+1}\right)^k$$

Also, $(O_1(|e_{\|}| + \|e_{\perp}\|))_k$ can be bounded by

$$\begin{aligned} (O_1(|e_{\|}| + \|e_{\perp}\|))_k &\leq \epsilon(|e_{\|}| + \|e_{\perp}\|))_k = \left(|e_{\|}|\epsilon\left(1 + \frac{\|e_{\perp}\|}{|e_{\|}|}\right)\right)_k \\ &= ((e_{\|})_k \epsilon)(|\widetilde{e}_{\|}| + \|\widetilde{e}_{\perp}\|)_0 \end{aligned}$$

so we can introduce $\widetilde{O}_1(\cdot)$ which is just a scaled version of $O_1(\cdot)$ with a smaller Lipschitz constant

$$\widetilde{\epsilon} \leq (e_{\|})_k \epsilon \leq \left(\frac{2}{1+m}\right)^k \epsilon$$

Similarly we can introduce $\widetilde{O}_2(\cdot)$ with a Lipschitz constant $\widetilde{M} \leq (e_\parallel)_k M$, writing $\left(O_2\left(\frac{\|e_\perp\|^2}{e_\parallel}\right)\right)_k \leq (e_\parallel)_k M \frac{\|\widetilde{e}_\perp\|^2}{\widetilde{e}_\parallel}$.

Observing that e_{k+1} corresponds to $\widetilde{e}_1$ under the same recursion we can apply equivalent of 3.4.0.7 (with new Lipschitz constants $\widetilde{\epsilon}$ and $\widetilde{M}$ for $\widetilde{O}_1(\cdot)$, $\widetilde{O}_2(\cdot)$) to $\widetilde{e}_0$ to get the following estimates

$$\begin{aligned}
\left\|\left(\frac{e_\perp}{e_\parallel}\right)_{k+1}\right\| &= \left\|\left(\frac{\widetilde{e}_\perp}{\widetilde{e}_\parallel}\right)_1\right\| \\
&\leq 2(2m\epsilon + m\eta M\sqrt{n-m})\left(\frac{2}{1+m}\right)^k \left(\frac{2}{1+m}\right)^{1+2+\cdots+k}\eta \\
&\leq \left(\frac{2}{1+m}\right)^{1+2+\cdots+k+(k+1)}\eta
\end{aligned}$$

thus giving us 3.4.0.5. Finally, 3.4.0.4 follows trivially from the way e_{k+1} is defined and 3.4.0.7, 3.4.0.8 applied to $\widetilde{e}_0$.

What can be said about $y_i = y(e_i)$? Recall 3.4.0.1

$$y(e) = \left(\frac{1}{m} + O_1\left(|e_\parallel| + \|e_\perp\|\right)\right)(-e_\perp) + V$$

and thus

$$\begin{aligned}
\|y(e)\| &\leq \left|\left(\frac{1}{m} + O_1\left(|e_\parallel| + \|e_\perp\|\right)\right)\right| \|(-e_\perp)\| + \|\mathbf{1}\|\left|O_2\left(\frac{\|e_\perp\|^2}{e_\parallel}\right)\right| \\
&\leq \left(\frac{1}{m} + 2\epsilon\right)\|e_\perp\| + \sqrt{n-m}M\eta\|e_\perp\|
\end{aligned}$$

(assuming $e_\parallel \leq 1$, $\frac{\|e_\perp\|}{e_\parallel} \leq \eta < 1$). Furthermore, if 3.4.0.10 is true, then

$$2\epsilon + \sqrt{n-m}M\eta \leq \frac{1}{m(1+m)}$$

and consequently

$$\begin{aligned}
\|y(e_i)\| &\leq \left(\tfrac{1}{m} + \tfrac{1}{m(1+m)}\right)\|(e_\perp)_i\| \\
&\leq \tfrac{2}{m}\|(e_\perp)_i\| \leq \tfrac{2}{m}\eta\left(\tfrac{2}{1+m}\right)^{1+2+\ldots+(i+1)-1} \qquad (3.4.0.14) \\
&= \tfrac{2}{m}\eta\left(\tfrac{2}{1+m}\right)^{\frac{(i+1)(i+2)}{2}-1}
\end{aligned}$$

thus giving us R-superlinear convergence rate for $\{y_i\}_{i\geq 0}$ (assuming $e_0 \in W_{\widehat{e}_0,\epsilon}$, a properly chosen wedge).

3.5 Concluding remarks and future research directions

The analysis above indicates an obvious direction for improvement of the convergence rate for $\{y_i\}_{i\geq 0}$. If one looks at the construction above carefully, we note that the main reason for getting just the (R-)superlinear rate of the order $\sim \varepsilon^{i^2}$ $(0 < \varepsilon < 1)$, but not the quadratic rate, is that there is a persisting non-quadratically diminishing error factor in the multiplier $\left(\frac{1}{m} + O_1(|e_{\|}| + \|e_{\perp}\|)\right)$ that we cannot eliminate, due to the relatively slow decay of $|e_{\|}|$ component (even if $\|e_{\perp}\|$ decreases very fast). So one possible remedy for this is to tackle this "slow" decay in $|e_{\|}|$ separately, for example, by making a two-step variant of the same algorithm where in the first step we target the decrease in $|e_{\|}|$ only (while possibly sacrificing $\|e_{\perp}\|$) and only in the second step aiming at the central line $\mathbb{L}_{M^c}$ (i.e., targeting the maximum possible decrease in $\|e_{\perp}\|$).

The preliminary study of this approach indeed shows that we can get a better convergence rate for the resulting $\{y_i\}_{i\geq 0}$ (e.g., R-quadratic rate and possibly even Q-quadratic), but there is much more work to be done in making this alternative algorithm "implementable", since, as of now, it is not quite clear when exactly one should switch to this two-step procedure (although we have an idea of how to make this switch happen automatically). On the contrary, the variant of the algorithm above would simply require that our algorithm behaves in the specified manner only asymptotically as we approach x^* (in a way, this is a "brain-dead" version of the algorithm, that would require you to execute this secant procedure only as $i \to \infty$).

Another direction for further analysis is eliminating the underlying assumption that $y(e_{i+1})$ is known precisely, while replacing it with, say, a Newton-like approximation from the current point (e_i, y_i).

Finally, to complete the analysis of this new optimization framework for LP, one needs to gain the full understanding of how to follow the path $\{(e(t), y(t)), t \geq 0\}$ outside of the wedge $W_{\widehat{e}_0,\epsilon}$ in the discrete setting, together with developing the resulting complexity estimates. Once this is done, the next step is to extend this framework for other hyperbolic programming problems, e.g., SDP.

Appendix A

Some linear algebra

We demonstrate how one can find inverses for two particularly structured matrices that we rely on in our analysis.

A.1 First matrix inverse

Suppose we want to find the inverse (if it exists) of

$$\begin{pmatrix} A & x \\ y^T & b \end{pmatrix}$$

where $A \in \mathbb{R}^{n \times n}$ is nonsingular, $x, y \in \mathbb{R}^n$, $b \in \mathbb{R}$.

To do this, we will be solving the system of linear equations of the form

$$\begin{pmatrix} A & x \\ y^T & b \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

for $u \in \mathbb{R}^n, v \in \mathbb{R}$, with $\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}$. We can rewrite this as

$$\begin{pmatrix} A^{-1} & \\ & 1 \end{pmatrix} \begin{pmatrix} A & x \\ y^T & b \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} A^{-1} & \\ & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

that is

$$\begin{pmatrix} I & A^{-1}x \\ y^T & b \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} A^{-1} & \\ & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

and furthermore

$$\begin{pmatrix} I & 0 \\ -y^T & 1 \end{pmatrix} \begin{pmatrix} I & A^{-1}x \\ y^T & b \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} I & 0 \\ -y^T & 1 \end{pmatrix} \begin{pmatrix} A^{-1} & \\ & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

giving us

$$\begin{pmatrix} I & A^{-1}x \\ 0 & -y^TA^{-1}x+b \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} I & 0 \\ -y^T & 1 \end{pmatrix}\begin{pmatrix} A^{-1} & \\ & 1 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

Proceeding with elimination of the block above the diagonal, we get

$$\begin{pmatrix} I(b-y^TA^{-1}x) & A^{-1}x \\ 0 & 1 \end{pmatrix}\begin{pmatrix} I & A^{-1}x \\ 0 & -y^TA^{-1}x+b \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}$$
$$= \begin{pmatrix} I(b-y^TA^{-1}x) & A^{-1}x \\ 0 & 1 \end{pmatrix}\begin{pmatrix} I & 0 \\ -y^T & 1 \end{pmatrix}\begin{pmatrix} A^{-1} & \\ & 1 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

which can be rewritten as

$$\begin{pmatrix} I(b-y^TA^{-1}x) & \\ & b-y^TA^{-1}x \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}$$
$$= \begin{pmatrix} I(b-y^TA^{-1}x) & A^{-1}x \\ 0 & 1 \end{pmatrix}\begin{pmatrix} I & 0 \\ -y^T & 1 \end{pmatrix}\begin{pmatrix} A^{-1} & \\ & 1 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

Finally

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} I\frac{1}{(b-y^TA^{-1}x)} & \\ & \frac{1}{b-y^TA^{-1}x} \end{pmatrix}$$
$$\begin{pmatrix} I(b-y^TA^{-1}x) & A^{-1}x \\ 0 & 1 \end{pmatrix}\begin{pmatrix} I & 0 \\ -y^T & 1 \end{pmatrix}\begin{pmatrix} A^{-1} & \\ & 1 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

and thus, the inverse (if it exists) must be

$$\begin{aligned}
\begin{pmatrix} A & x \\ y^T & b \end{pmatrix}^{-1} &= \begin{pmatrix} I\frac{1}{(b-y^TA^{-1}x)} & \\ & \frac{1}{b-y^TA^{-1}x} \end{pmatrix} \\
&\quad \begin{pmatrix} I(b-y^TA^{-1}x) & A^{-1}x \\ 0 & 1 \end{pmatrix} \begin{pmatrix} I & 0 \\ -y^T & 1 \end{pmatrix} \begin{pmatrix} A^{-1} & \\ & 1 \end{pmatrix} \\
&= \frac{1}{b-y^TA^{-1}x} \begin{pmatrix} A^{-1}(b-y^TA^{-1}x)+(A^{-1}x)(y^TA^{-1}) & -A^{-1}x \\ -y^TA^{-1} & 1 \end{pmatrix} \\
&= \begin{pmatrix} A^{-1} & \\ & 0 \end{pmatrix} + \frac{1}{b-y^TA^{-1}x} \begin{pmatrix} -A^{-1}x \\ 1 \end{pmatrix} \begin{pmatrix} -(A^T)^{-1}y \\ 1 \end{pmatrix}^T
\end{aligned}$$

a rank-1 perturbation of $\begin{pmatrix} A^{-1} & \\ & 0 \end{pmatrix}$. The existence of the inverse is easy to check now: we need to make sure that $b - y^TA^{-1}x \neq 0$.

A.2 Second matrix inverse

Suppose we are given a non-singular $A \in \mathbb{R}^{(n-m)\times(n-m)}$, $B \in \mathbb{R}^{((n-m)-1)\times(n-m)}$ is defined as

$$B := \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ 1 & 0 & -1 & 0 & \dots & 0 \\ 1 & 0 & 0 & -1 & \dots & 0 \\ & & \dots & & & \\ 1 & 0 & 0 & 0 & \dots & -1 \end{bmatrix}$$

and a vector $y \in \mathbb{R}^{n-m}$. We want to find the inverse

$$\begin{pmatrix} BA \\ y^T \end{pmatrix}$$

denoted by X. Write

$$X \begin{pmatrix} BA \\ y^T \end{pmatrix} = I \quad \Leftrightarrow \quad X \begin{pmatrix} B \\ y^T A^{-1} \end{pmatrix} = A^{-1}$$

so we need to invert

$$\begin{pmatrix} B \\ y^T A^{-1} \end{pmatrix} = \begin{pmatrix} B \\ z^T \end{pmatrix}$$

with $z^T = y^T A^{-1}$. This matrix has a special structure, namely, if we let

$$T := \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ 1 & 0 & -1 & 0 & \dots & 0 \\ 1 & 0 & 0 & -1 & \dots & 0 \\ & & \dots & & & \\ 1 & 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

then it can be written as

$$\begin{pmatrix} B \\ z^T \end{pmatrix} = \left[\begin{array}{c|c} & \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix} \\ T & \\ & \\ \hline z_{-k}^T & z_k \end{array}\right]$$

with $z \in \mathbb{R}^{n-m} \equiv \mathbb{R}^k$, $z_{-k} = (z_1, z_2, \dots, z_{k-1}) \in \mathbb{R}^{k-1}$ (and $T \in \mathbb{R}^{(k-1)\times(k-1)}$). Note that T is invertible,

$$T^{-1} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 1 \\ -1 & 0 & 0 & \dots & 0 & 1 \\ 0 & -1 & 0 & \dots & 0 & 1 \\ & & \dots & & & \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}$$

Introducing

$$-e_{k-1}^T = (0, 0, \ldots, 0, -1)$$

we can write (see A.1 above)

$$\begin{aligned}
\begin{pmatrix} B \\ z^T \end{pmatrix}^{-1} &= \begin{bmatrix} T^{-1} & 0 \\ 0 & 0 \end{bmatrix} \\
&+ \frac{1}{z_k - z_{-k}^T T^{-1}(-e_{k-1})} \begin{bmatrix} T^{-1}(-e_{k-1})(z_{-k}^T T^{-1}) & -T^{-1}(-e_{k-1}) \\ -z_{-k}^T T^{-1} & 1 \end{bmatrix} \\
&= \begin{bmatrix} T^{-1} & 0 \\ 0 & 0 \end{bmatrix} \\
&+ \frac{1}{(\mathbf{1}^T z)} \begin{bmatrix} \begin{pmatrix} -1 \\ -1 \\ -1 \\ \vdots \\ -1 \end{pmatrix} (-z_2, -z_3, -z_4, \ldots, -z_{k-1}, \mathbf{1}^T z_{-k}) & \begin{pmatrix} -1 \\ -1 \\ -1 \\ \vdots \\ -1 \end{pmatrix} \\ (z_2, z_3, z_4, \ldots, z_{k-1}, -\mathbf{1}^T z_{-k}) & 1 \end{bmatrix} \\
&= \begin{bmatrix} T^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{(\mathbf{1}^T z)} \begin{bmatrix} z_2 & z_3 & z_4 & \ldots & z_{k-1} & -\mathbf{1}^T z_{-k} & 1 \\ z_2 & z_3 & z_4 & \ldots & z_{k-1} & -\mathbf{1}^T z_{-k} & 1 \\ & & & \vdots & & & \\ z_2 & z_3 & z_4 & \ldots & z_{k-1} & -\mathbf{1}^T z_{-k} & 1 \end{bmatrix} \\
&= \begin{bmatrix} T^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{(\mathbf{1}^T z)} \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} (z_2, z_3, z_4, \ldots, z_{k-1}, -\mathbf{1}^T z_{-k}, 1)
\end{aligned}$$

Therefore

$$\begin{pmatrix} BA \\ y^T \end{pmatrix}^{-1} = A^{-1} \begin{pmatrix} B \\ z^T \end{pmatrix}^{-1}$$

$$= A^{-1} \left(\begin{bmatrix} T^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{(\mathbf{1}^T z)} \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} (z_2, z_3, \ldots, z_{k-1}, -\mathbf{1}^T z_{-k}, 1) \right)$$

where $z^T = y^T A^{-1}$ (a rank-one perturbation of the product of A^{-1} and T^{-1} block matrix). The condition for the existence of the inverse is now obvious: $\mathbf{1}^T z \neq 0$.

Appendix B

Essential results for Newton's method complexity analysis

Here we quote the main results borrowed from [19]. For the proofs and the complete exposition of the material consult the monograph itself.

Let $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ be C^1. We are concerned with finding a root of f, that is, solving the equation

$$f(z) = 0$$

If at a point z_0 the derivative of f is non-singular, we can solve a similar equation using linearization of f at that point

$$f(z_0) + f'(z_0)dz = 0$$

If the first order Taylor expansion of f at a current point z_0 is "accurate enough", we would hope that the true root of f can be well approximated by $z_1 := z_0 + dz$. Furthermore, we can build such a linearization of f at z_1 and repeat this procedure to find z_2, and so on. This is Newton's method (an iterative procedure for finding a root of a C^1 function).

Newton's method is an iteration based on the map from $\mathbb{R}^n$ to itself,

$$N_f(z) = z - (f'(z))^{-1} f(z)$$

(this formula is defined as long as $(f'(z))^{-1}$ exists).

We will denote the Newton step for f as

$$NS_f(z) := -(f'(z))^{-1} f(z)$$

Definition B.0.1 (Definition 1 in [19]). Say that z is an *approximate zero* of f if the sequence given by $z_0 = z$ and $z_{i+1} = N_f(z_i)$ is defined for all natural numbers i, and there is a ξ such that $f(\xi) = 0$ with

$$\|z_i - \xi\| \leq \left(\frac{1}{2}\right)^{2^i - 1} \|z - \xi\|$$

Call ξ the *associated zero.*

Define an auxiliary quantity

$$\gamma = \gamma(f, z) = \sup_{k \geq 2} \left\| \frac{f'(z)^{-1} f^{(k)}(z)}{k!} \right\|^{1/k-1}$$

where $f^{(k)}$ is the k^{th} derivative of f. This definition applies to analytic functions f. If f is (real) analytic and $f(z)^{-1}$ exists, then the sup exists as well since $f^{(k)}/k! = a_k$ has a geometric growth rate.

Theorem B.0.2 (Theorem 1 in [19]). *Suppose that $f(\xi) = 0$ and that $f'(\xi)^{-1}$ exists. If*

$$\|z - \xi\| \leq \frac{3 - \sqrt{7}}{2\gamma} \text{ for } \gamma = \gamma(f, \xi)$$

then z is an approximate zero of f with associated zero ξ.

Note: this implies uniqueness of ξ.

The following simple polynomial plays an important role in the estimates presented

$$\psi(u) = 1 - 4u + 2u^2$$

Proposition B.0.3 (Proposition 1 in [19]). *Let $f(\xi) = 0$, and let $u = \|z - \xi\|\gamma(f, \xi)$. Suppose $u < (5 - \sqrt{17})/4$. Then*

(a) $\|N_f(z) - \xi\| < \frac{\gamma(f,\xi)\|z-\xi\|^2}{\psi(u)} = \frac{u\|z-\xi\|}{\psi(u)}$

(b) $\|N_f^k(z) - \xi\| \leq \left(\frac{u}{\psi(u)}\right)^{2^k-1} \|z - \xi\|$ *for all* $k \geq 0$ *(where* $N_f^k(z)$ *is the* k^{th} *Newton's iterate starting from* z*)*

Define two more auxiliary quantities, the length of the Newton step $NS_f(z)$

$$\beta(f, z) = \|z - N_f(z)\| = \|NS_f(z)\| = \|f'(z)^{-1} f(z)\|$$

and

$$\alpha(f, z) = \beta(f, z)\gamma(f, z)$$

Theorem B.0.4 (Theorem 2 in [19])**.** *There is a universal constant* α_0 *with the following property. If* $\alpha(f, z) < \alpha_0$*, then* z *is an approximate zero of* f *in the sense of Definition B.0.1. Moreover, the distance form* z *to the associated zero* ξ *is at most* $2\beta(f, z)$*.*

Remark B.0.5 (Remark 1 in [19])*.* The invariant $\alpha(f, z)$ depends only on derivatives of f at the point z, which can be computed if f is a polynomial map. Thus Theorem B.0.4 gives a criterion that can be used in principle and in practice to give certainty that z is indeed an approximation to a solution.

In particular α_0 can be chosen to be .03.

Proposition B.0.6 (Proposition 3 in [19])**.** *If* $u < 1-(\sqrt{2}/2)$ *and* $\|z_1-z\|\gamma(f, z) = u$*, then*

(a) $\beta(f, z_1) \leq \frac{(1-u)}{\psi(u)}((1 - u)\beta(f, z) + \|z_1 - z\|)$*;*

(b) $\gamma(f, z_1) \leq \frac{\gamma(f,z)}{\psi(u)(1-u)}$*;*

(c) $\alpha(f, z_1) \leq \frac{(1-u)\alpha(f,z)+u}{\psi(u)^2}$

Proposition B.0.7 (Proposition 6 in [19])**.** *Let* f *be analytic at* z *and* r *be the radius of convergence of the Taylor series of* f *at* z*. Then* $r \geq 1/\gamma(f, z)$*.*

Appendix C

Proof of Corollary 3.3.15

We present the proof by induction (on the dimension of $\mathbb{R}^k$).

Since $\widetilde{f}$ is C^1 at $(\rho, \phi) = (0, \mathbf{0})$, it is enough to compute a directional derivative as $(\rho, \phi) \to (0, \mathbf{0})$ using f, in some particular direction (we will chose $\phi = \mathbf{0}, \rho \downarrow 0$) to evaluate $\widetilde{f}$ and its Jacobian at the origin.

Induction base: let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be C^1

$$f(x_1, x_2) = \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix}$$

Assuming the matrix

$$\begin{pmatrix} \frac{\partial x_1}{\partial \rho} & \frac{\partial x_1}{\partial \phi} \\ \frac{\partial x_2}{\partial \rho} & \frac{\partial x_2}{\partial \phi} \end{pmatrix} = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} 1 & \\ & \rho \end{pmatrix}$$

is invertible, we write

$$\begin{aligned} \widetilde{(}\rho, \phi) &= \begin{pmatrix} \frac{\partial x_1}{\partial \rho} & \frac{\partial x_1}{\partial \phi} \\ \frac{\partial x_2}{\partial \rho} & \frac{\partial x_2}{\partial \phi} \end{pmatrix}^{-1} \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix} \\ &= \begin{pmatrix} 1 & \\ & \frac{1}{\rho} \end{pmatrix} \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix} \\ &= \begin{pmatrix} f_1(x_1, x_2)\cos\phi + f_2(x_1, x_2)\sin\phi \\ \frac{1}{\rho}(-f_1(x_1, x_2)\sin\phi + f_2(x_1, x_2)\cos\phi) \end{pmatrix} \end{aligned}$$

As $\rho \to 0$ with $\phi = 0$

$$\widetilde{f_1}(\rho, \phi) = f_1(x_1, x_2)\cos\phi + f_2(x_1, x_2)\sin\phi \to 0$$

since $(x_1, x_2) \to \mathbf{0}$ and thus $f(x_1, x_2) \to \mathbf{0}$, also

$$\begin{aligned}
\widetilde{f}_2(\rho,\phi) &= \frac{1}{\rho}(-f_1(x_1,x_2)\sin\phi + f_2(x_1,x_2)\cos\phi) \\
&\to \frac{\left(-\frac{\partial}{\partial\rho}f_1(x_1,x_2)\sin\phi + \frac{\partial}{\partial\rho}f_2(x_1,x_2)\cos\phi\right)}{1} \\
&= -\left(\frac{\partial}{\partial x_1}f_1(x_1,x_2)\frac{\partial x_1}{\partial\rho} + \frac{\partial}{\partial x_2}f_1(x_1,x_2)\frac{\partial x_2}{\partial\rho}\right)\sin\phi \\
&+ \left(\frac{\partial}{\partial x_1}f_2(x_1,x_2)\frac{\partial x_1}{\partial\rho} + \frac{\partial}{\partial x_2}f_2(x_1,x_2)\frac{\partial x_2}{\partial\rho}\right)\cos\phi \\
&= -\left(\frac{\partial}{\partial x_1}f_1(x_1,x_2)\cos\phi + \frac{\partial}{\partial x_2}f_1(x_1,x_2)\sin\phi\right)\sin\phi \\
&+ \left(\frac{\partial}{\partial x_1}f_2(x_1,x_2)\cos\phi + \frac{\partial}{\partial x_2}f_2(x_1,x_2)\sin\phi\right)\cos\phi \\
&= 0
\end{aligned}$$

since the Jacobian of f is assumed to have diagonal structure.

To evaluate the Jacobian for $\widetilde{f}$ at the origin we write

$$\widetilde{f}(\rho,\phi) = \begin{pmatrix} \frac{\partial x_1}{\partial\rho} & \frac{\partial x_1}{\partial\phi} \\ \frac{\partial x_2}{\partial\rho} & \frac{\partial x_2}{\partial\phi} \end{pmatrix}^{-1} \begin{pmatrix} f_1(x_1,x_2) \\ f_2(x_1,x_2) \end{pmatrix}$$

with

$$\begin{pmatrix} \frac{\partial x_1}{\partial\rho} & \frac{\partial x_1}{\partial\phi} \\ \frac{\partial x_2}{\partial\rho} & \frac{\partial x_2}{\partial\phi} \end{pmatrix}^{-1} = \begin{pmatrix} 1 & \\ & \frac{1}{\rho} \end{pmatrix} \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}$$

so that

$$\nabla_{\rho,\phi}\widetilde{f} = \nabla_{\rho,\phi}\left(\left(\begin{pmatrix} \frac{\partial x_1}{\partial\rho} & \frac{\partial x_1}{\partial\phi} \\ \frac{\partial x_2}{\partial\rho} & \frac{\partial x_2}{\partial\phi} \end{pmatrix}^{-1}\right)\begin{pmatrix} f_1 \\ f_2 \end{pmatrix} + \begin{pmatrix} \frac{\partial x_1}{\partial\rho} & \frac{\partial x_1}{\partial\phi} \\ \frac{\partial x_2}{\partial\rho} & \frac{\partial x_2}{\partial\phi} \end{pmatrix}^{-1} \nabla_{\rho,\phi}\begin{pmatrix} f_1 \\ f_2 \end{pmatrix}\right)$$

(by the chain rule). The first term in this expression is given by

$$\begin{aligned}
\frac{\partial}{\partial\rho}\left(\left(\begin{pmatrix} \frac{\partial x_1}{\partial\rho} & \frac{\partial x_1}{\partial\phi} \\ \frac{\partial x_2}{\partial\rho} & \frac{\partial x_2}{\partial\phi} \end{pmatrix}^{-1}\right)\right) &= \frac{\partial}{\partial\rho}\left(\left(\begin{pmatrix} 1 & \\ & \frac{1}{\rho} \end{pmatrix}\begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}\right)\right) \\
&= \begin{pmatrix} 0 & 0 \\ \frac{\sin\phi}{\rho^2} & \frac{-\cos\phi}{\rho^2} \end{pmatrix}
\end{aligned}$$

and

$$\frac{\partial}{\partial\phi}\left(\begin{pmatrix} \frac{\partial x_1}{\partial\rho} & \frac{\partial x_1}{\partial\phi} \\ \frac{\partial x_2}{\partial\rho} & \frac{\partial x_2}{\partial\phi} \end{pmatrix}^{-1}\right) = \begin{pmatrix} -\sin\phi & \cos\phi \\ \frac{-\cos\phi}{\rho} & \frac{-\sin\phi}{\rho} \end{pmatrix}$$

post-multiplied by $\begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$ (giving us first and second columns of the Jacobian respectively). The second term is given by

$$\begin{pmatrix} \frac{\partial x_1}{\partial\rho} & \frac{\partial x_1}{\partial\phi} \\ \frac{\partial x_2}{\partial\rho} & \frac{\partial x_2}{\partial\phi} \end{pmatrix}^{-1} \nabla_{\rho,\phi}\begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = \begin{pmatrix} \frac{\partial x_1}{\partial\rho} & \frac{\partial x_1}{\partial\phi} \\ \frac{\partial x_2}{\partial\rho} & \frac{\partial x_2}{\partial\phi} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial f_1}{\partial\rho} & \frac{\partial f_1}{\partial\phi} \\ \frac{\partial f_2}{\partial\rho} & \frac{\partial f_2}{\partial\phi} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial x_1}{\partial\rho} & \frac{\partial x_1}{\partial\phi} \\ \frac{\partial x_2}{\partial\rho} & \frac{\partial x_2}{\partial\phi} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{pmatrix} \begin{pmatrix} \frac{\partial x_1}{\partial\rho} & \frac{\partial x_1}{\partial\phi} \\ \frac{\partial x_2}{\partial\rho} & \frac{\partial x_2}{\partial\phi} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \\ & \frac{1}{\rho} \end{pmatrix} \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{pmatrix} \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} 1 & \\ & \rho \end{pmatrix}$$

We have

$$\begin{aligned} \nabla_{\rho,\phi}\widetilde{f} &= \begin{pmatrix} 0 & -f_1\sin\phi + f_2\cos\phi \\ \frac{1}{\rho^2}(f_1\sin\phi - f_2\cos\phi) & \frac{1}{\rho}(-f_1\cos\phi - f_2\sin\phi) \end{pmatrix} \\ &+ \begin{pmatrix} \cos\phi\frac{\partial f_1}{\partial\rho} + \sin\phi\frac{\partial f_2}{\partial\rho} & \cos\phi\frac{\partial f_1}{\partial\phi} + \sin\phi\frac{\partial f_2}{\partial\phi} \\ \frac{1}{\rho}(-\sin\phi\frac{\partial f_1}{\partial\rho} + \cos\phi\frac{\partial f_2}{\partial\rho}) & \frac{1}{\rho}(-\sin\phi\frac{\partial f_1}{\partial\phi} + \cos\phi\frac{\partial f_2}{\partial\phi}) \end{pmatrix} \end{aligned}$$

and we are interested in the limit of this expression with $\phi = 0$ as $\rho \to 0$. Evaluating each term (recalling that f is assumed to have diagonal Jacobian now) gives us

$$\begin{aligned} \left(\nabla_{\rho,\phi}\widetilde{f}\right)_{11} &= \cos\phi\frac{\partial f_1}{\partial\rho} + \sin\phi\frac{\partial f_2}{\partial\rho} \\ &= \cos\phi\left(\frac{\partial f_1}{\partial x_1}\cos\phi + \frac{\partial f_1}{\partial x_2}\sin\phi\right) \\ &+ \sin\phi\left(\frac{\partial f_2}{\partial x_1}\cos\phi + \frac{\partial f_2}{\partial x_2}\sin\phi\right) \to \frac{\partial f_1}{\partial x_1} \end{aligned}$$

$$\begin{aligned}\left(\nabla_{\rho,\phi}\widetilde{f}\right)_{12} &= \left(-f_1+\frac{\partial f_2}{\partial\phi}\right)\sin\phi+\left(f_2+\frac{\partial f_1}{\partial\phi}\right)\cos\phi\\ &= -f_1\sin\phi+f_2\cos\phi\\ &+ \cos\phi\left(\frac{\partial f_1}{\partial x_1}(-\rho\sin\phi)+\frac{\partial f_1}{\partial x_2}\rho\cos\phi\right)\\ &+ \sin\phi\left(\frac{\partial f_2}{\partial x_1}(-\rho\sin\phi)+\frac{\partial f_2}{\partial x_2}\rho\cos\phi\right)\to 0\end{aligned}$$

$$\begin{aligned}\left(\nabla_{\rho,\phi}\widetilde{f}\right)_{21} &= \frac{1}{\rho^2}(f_1\sin\phi-f_2\cos\phi)+\frac{1}{\rho}\left(-\sin\phi\frac{\partial f_1}{\partial\rho}+\cos\phi\frac{\partial f_2}{\partial\rho}\right)\\ &= \frac{1}{\rho^2}(f_1\sin\phi-f_2\cos\phi)\\ &+ \frac{1}{\rho}\left(-\sin\phi\left(\frac{\partial f_1}{\partial x_1}\cos\phi+\frac{\partial f_1}{\partial x_2}\sin\phi\right)\right)\\ &+ \frac{1}{\rho}\left(\cos\phi\left(\frac{\partial f_2}{\partial x_1}\cos\phi+\frac{\partial f_2}{\partial x_2}\sin\phi\right)\right)\\ &= \frac{-f_2}{\rho^2}\to\frac{1}{2\rho}\frac{\partial f_2}{\partial\rho}=\frac{1}{2\rho}\left(\frac{\partial f_2}{\partial x_1}\cos\phi+\frac{\partial f_2}{\partial x_2}\sin\phi\right)\to 0\end{aligned}$$

and

$$\begin{aligned}\left(\nabla_{\rho,\phi}\widetilde{f}\right)_{22} &= \frac{1}{\rho}(-f_1\cos\phi-f_2\sin\phi)+\frac{1}{\rho}\left(-\sin\phi\frac{\partial f_1}{\partial\phi}+\cos\phi\frac{\partial f_2}{\partial\phi}\right)\\ &= \frac{1}{\rho}(-f_1\cos\phi)+\frac{1}{\rho}\cos\phi\left(\frac{\partial f_2}{\partial x_1}(-\rho\sin\phi)+\frac{\partial f_2}{\partial x_2}\rho\cos\phi\right)\\ &= \frac{1}{\rho}(-f_1)+\frac{\partial f_2}{\partial x_2}\to-\frac{\partial f_1}{\partial\rho}+\frac{\partial f_2}{\partial x_2}\\ &= \frac{\partial f_2}{\partial x_2}-\left(\frac{\partial f_1}{\partial x_1}\cos\phi+\frac{\partial f_1}{\partial x_2}\sin\phi\right)\\ &= \frac{\partial f_2}{\partial x_2}-\frac{\partial f_1}{\partial x_1}\end{aligned}$$

So as $\rho\to 0$ with $\phi=0$ we have

$$\nabla_{\rho,\phi}\widetilde{f}\to\begin{pmatrix}\frac{\partial f_1}{\partial x_1} & \\ & \left(\frac{\partial f_2}{\partial x_2}-\frac{\partial f_1}{\partial x_1}\right)\end{pmatrix}$$

Inductive step: assume it is true in $\mathbb{R}^k$, we want to show it is true in $\mathbb{R}^{k+1}$.

Consider $\widetilde{f}$ in $\mathbb{R}^n$. Note that at $\phi = \mathbf{0}$

$$\nabla_{\rho,\phi}\Psi_n(\rho,\phi) = \begin{pmatrix} 1 & & & & \\ & \rho & & & \\ & & \rho & & \\ & & & \ddots & \\ & & & & \rho \end{pmatrix}$$

so that

$$\nabla_{\rho,\phi}\Psi_n(\rho,\phi)^{-1} = \begin{pmatrix} 1 & & & & \\ & \frac{1}{\rho} & & & \\ & & \frac{1}{\rho} & & \\ & & & \ddots & \\ & & & & \frac{1}{\rho} \end{pmatrix}$$

and therefore with $\phi = \mathbf{0}$

$$\lim_{\rho\downarrow 0} \nabla_{\rho,\phi}\Psi_n^{-1} f(\Psi(\rho,\phi)) = 0$$

since for $i \geq 2$

$$\lim_{\rho\downarrow 0} \frac{f_i(\Psi(\rho,\phi))}{\rho} = \lim_{\rho\downarrow 0} \frac{\sum_{j=1}^n \frac{\partial f_i}{\partial x_j}\frac{\partial x_j}{\rho}}{1} = \lim_{\rho\downarrow 0} \frac{\frac{\partial f_i}{\partial x_i}\frac{\partial x_i}{\rho}}{1} = 0$$

So $\widetilde{f}(0,\mathbf{0}) = (0,\mathbf{0})$.

How can we compute the linearization of $\widetilde{f}$ at $(\rho,\phi) = (0,\mathbf{0})$? Observe that the Jacobian of $\widetilde{f}$ (the matrix whose first column is the derivative of $\widetilde{f}$ with respect to ρ, second – with respect to ϕ_1, etc.) satisfies

$$\nabla_{\rho,\phi}\widetilde{f}(\rho,\phi) = \nabla_{\rho,\phi}\left[\nabla_{\rho,\phi}\Psi_n(\rho,\phi)^{-1}\right] f(\Psi_n(\rho,\phi)) + \nabla_{\rho,\phi}\Psi_n(\rho,\phi)^{-1}\nabla_{\rho,\phi} f(\Psi_n(\rho,\phi))$$

At $\phi = \mathbf{0}$, the second term in this expression, $\nabla_{\rho,\phi}\Psi_n(\rho,\phi)^{-1}\nabla_{\rho,\phi} f(\Psi_n(\rho,\phi))$,

(by the chain rule) is

$$\begin{pmatrix} 1 & & & \\ & \frac{1}{\rho} & & \\ & & \frac{1}{\rho} & \\ & & & \ddots \end{pmatrix} \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} & \cdots \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_2}{\partial x_3} & \cdots \\ \frac{\partial f_3}{\partial x_1} & \frac{\partial f_3}{\partial x_2} & \frac{\partial f_3}{\partial x_3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \frac{\partial x_1}{\partial \rho} & \frac{\partial x_1}{\partial \phi_1} & \frac{\partial x_1}{\partial \phi_2} & \cdots \\ \frac{\partial x_2}{\partial \rho} & \frac{\partial x_2}{\partial \phi_1} & \frac{\partial x_2}{\partial \phi_2} & \cdots \\ \frac{\partial x_3}{\partial \rho} & \frac{\partial x_3}{\partial \phi_1} & \frac{\partial x_3}{\partial \phi_2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$= \begin{pmatrix} 1 & & & \\ & \frac{1}{\rho} & & \\ & & \frac{1}{\rho} & \\ & & & \ddots \end{pmatrix} \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & & & \\ & \frac{\partial f_2}{\partial x_2} & & \\ & & \frac{\partial f_3}{\partial x_3} & \\ & & & \ddots \end{pmatrix} \begin{pmatrix} 1 & & & \\ & \rho & & \\ & & \rho & \\ & & & \ddots \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & & & \\ & \frac{\partial f_2}{\partial x_2} & & \\ & & \frac{\partial f_3}{\partial x_3} & \\ & & & \ddots \end{pmatrix}$$

and this is true in the limit as $\rho \to 0$ as well.

The evaluation of the first term in the expression for the Jacobian of $\widetilde{f}$ is more involved and requires the knowledge of partial derivatives of $\nabla_{\rho,\phi}\Psi_n^{-1}$. To make the notation more compact, we consider $f \in \mathbb{R}^{n+1}$ instead.

Observe

$$\nabla_{\rho,\phi}\Psi_{n+1} = \left[\begin{array}{c|c} (\nabla_{\rho,\phi_{-(n-1)}}\Psi_{n+1})\cos\phi_n & \begin{pmatrix} -(((\rho\cos\phi_1)\cos\phi_2)\cdots\cos\phi_{n-1})\sin\phi_n \\ -(((\rho\sin\phi_1)\cos\phi_2)\cdots\cos\phi_{n-1})\sin\phi_n \\ \vdots \\ -(\rho\sin\phi_{n-1})\sin\phi_n \end{pmatrix} \\ \hline (\sin\phi_n, 0, 0, \ldots, 0) & \rho\cos\phi_n \end{array}\right]$$

has a recursive structure. Denoting

$$A = (\nabla_{\rho,\phi_{-(n-1)}}\Psi_{n+1})\cos\phi_n$$

$$x = \begin{pmatrix} -(((\rho\cos\phi_1)\cos\phi_2)\cdots\cos\phi_{n-1})\sin\phi_n \\ -(((\rho\sin\phi_1)\cos\phi_2)\cdots\cos\phi_{n-1})\sin\phi_n \\ \vdots \\ -(\rho\sin\phi_{n-1})\sin\phi_n \end{pmatrix}$$

$$y^T = (\sin\phi_n, 0, 0, \ldots, 0)$$

and

$$b = \rho\cos\phi_n$$

we have

$$\begin{aligned} (\nabla_{\rho,\phi}\Psi_{n+1})^{-1} &= \begin{bmatrix} A & x \\ y^T & b \end{bmatrix}^{-1} \\ &= \begin{bmatrix} A^{-1} & 0 \\ 0 & 0 \end{bmatrix}^{-1} + \frac{1}{b - y^T A^{-1} x} \begin{pmatrix} -A^{-1}x \\ 1 \end{pmatrix} \begin{pmatrix} -(A^T)^{-1}y \\ 1 \end{pmatrix}^T \end{aligned}$$

(see A.1) assuming, of course, $b - y^T A^{-1} x \neq 0$ (which is true).

In order to complete this evaluation we need to compute $\frac{\partial}{\partial\rho}, \frac{\partial}{\partial\phi_{-n}}, \frac{\partial}{\partial\phi_n}$ of $(\nabla_{\rho\phi}\Psi_{n+1})^{-1}$. Consider the partial derivative with respect to ρ at $\phi = 0$:

$$A = A(\rho) = \begin{pmatrix} 1 & & & & \\ & \rho & & & \\ & & \rho & & \\ & & & \ddots & \\ & & & & \rho \end{pmatrix}$$

so that

$$A^{-1}(\rho) = \begin{pmatrix} 1 & & & & \\ & \frac{1}{\rho} & & & \\ & & \frac{1}{\rho} & & \\ & & & \ddots & \\ & & & & \frac{1}{\rho} \end{pmatrix}$$

Also

$$x^T = x^T(\rho) = -(0, 0, 0, \ldots, 0)$$

$$y^T = y^T(\rho) = (0, 0, 0, \ldots, 0)$$

$$b = b(\rho) = \rho$$

With this in mind

$$\begin{aligned} \frac{\partial}{\partial \rho}\left((\nabla_{\rho,\phi}\Psi_{n+1})^{-1}\right) &= \frac{\partial}{\partial \rho}\left(\begin{bmatrix} A^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{\rho}\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}^T\right) \\ &= \frac{\partial}{\partial \rho}\left(\begin{bmatrix} A^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ & & \vdots & & \\ 0 & 0 & \ldots & 0 & \frac{1}{\rho} \end{bmatrix}\right) \\ &= \begin{pmatrix} 0 & & & \\ & -\frac{1}{\rho^2} & & \\ & & \ddots & \\ & & & -\frac{1}{\rho^2} \end{pmatrix} \end{aligned}$$

so that the first column of the first summand in the expression for the Jacobian of $\widetilde{f}$ (corresponding to $\frac{\partial}{\partial\rho}$) will be

$$\lim_{\rho\downarrow 0}\begin{pmatrix} 0 & & & \\ & -\frac{1}{\rho^2} & & \\ & & \ddots & \\ & & & -\frac{1}{\rho^2} \end{pmatrix} f(\Psi_{n+1}(\rho,\phi)) = 0$$

since

$$\lim_{\rho\downarrow 0} -\frac{1}{\rho^2} f_{n+1}(\Psi_{n+1}(\rho,\phi)) = \lim_{\rho\downarrow 0} \frac{\sum_{i=1}^{n+1} \frac{\partial f_{n+1}}{\partial x_i}\frac{\partial x_i}{\partial\rho}}{-2\rho} = \lim_{\rho\downarrow 0} \frac{\frac{\partial f_{n+1}}{\partial x_{n+1}}\frac{\partial x_{n+1}}{\partial\rho}}{-2\rho} = 0$$

(all the other terms are evaluated similarly and are also 0; note that this is the same as in the case of Ψ_n).

Consider the partial derivative with respect to ϕ_{-n}: A^{-1}, as a function of ϕ_{-n} and ρ, $A^{-1}(\phi_{-n};\rho)$, has the same effect as for Ψ_n ("damping" the diagonal of the Jacobian) since the only difference is the multiplication by $\cos\phi_n = 1$ (since $\phi = 0$), which is independent of ϕ_{-n}. The corresponding rank-one update for the inverse has

$$b = b(\phi_{-n};\rho) = \rho$$

$$x^T = x^T(\phi_{-n};\rho) = \mathbf{0}$$

$$y^T = y^T(\phi_{-n};\rho) = \mathbf{0}$$

(we leave the dependence on ρ since we are interested in the limit as $\rho \downarrow 0$) and gives 0 as $\frac{\partial}{\partial\phi_{-n}}$ additional block. Thus, $\frac{\partial}{\partial\phi_{-n}}$ gives the same block in the Jacobian of $\widetilde{f}$ as in the case of Ψ_n ("damping" of the diagonal elements except for the first and the last ones).

Lastly, the $\frac{\partial}{\partial \phi_n}$ term is given by

$$A^{-1} = A^{-1}(\phi_n; \rho) = \begin{pmatrix} 1 & & & & \\ & \frac{1}{\rho} & & & \\ & & \frac{1}{\rho} & & \\ & & & \ddots & \\ & & & & \frac{1}{\rho} \end{pmatrix} \frac{1}{\cos\phi_n}$$

$$x^T = x^T(\phi_n; \rho) = -\rho(\sin\phi_n, 0, 0, \ldots, 0)$$

$$y^T = y^T(\phi_n; \rho) = (\sin\phi_n, 0, 0, \ldots, 0)$$

$$b = b(\phi_n; \rho) = \rho\cos\phi_n$$

so that

$$(A^{-1}x)^T = \frac{1}{\cos\phi_n}(-\rho)(\sin\phi_n, 0, 0, \ldots, 0) = \left(-\rho\frac{\sin\phi_n}{\cos\phi_n}, 0, 0, \ldots, 0\right)$$

$$y^T A^{-1} = \left(\frac{\sin\phi_n}{\cos\phi_n}, 0, 0, \ldots, 0\right)$$

$$y^T A^{-1} x = (\sin\phi_n, 0, 0, \ldots, 0)(-\rho)\begin{pmatrix} \frac{\sin\phi_n}{\cos\phi_n} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = -\rho\left(\frac{\sin^2\phi_n}{\cos\phi_n}\right)$$

and consequently

$$\frac{\partial}{\partial\phi_n}\left(\nabla_{\rho,\phi}\Psi_{n+1}^{-1}\right) = \frac{\partial}{\partial\phi_n}\left[\begin{array}{c|c} \frac{1}{\cos\phi_n}\begin{pmatrix} 1 & & & \\ & \frac{1}{\rho} & & \\ & & \frac{1}{\rho} & \\ & & & \ddots \end{pmatrix} & \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\ \hline (0,0,\ldots,0) & 0 \end{array}\right]$$

$$+ \frac{\partial}{\partial\phi_n}\left[\left(\frac{1}{\rho\cos\phi_n - \rho\frac{\sin^2\phi_n}{\cos\phi_n}}\right)\begin{pmatrix} \rho\frac{\sin\phi_n}{\cos\phi_n} \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}\begin{pmatrix} -\frac{\sin\phi_n}{\cos\phi_n} \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}^T\right]$$

$$= \left[\begin{array}{c|c} \frac{-\sin\phi_n}{\cos^2\phi_n}\begin{pmatrix} 1 & & & \\ & \frac{1}{\rho} & & \\ & & \frac{1}{\rho} & \\ & & & \ddots \end{pmatrix} & \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\ \hline (0,0,\ldots,0) & 0 \end{array}\right]$$

$$+ \frac{\partial}{\partial\phi_n}\left[\frac{1\cos^2\phi_n}{\rho\cos 2\phi_n}\frac{1}{\cos\phi_n}\begin{pmatrix} \rho\sin\phi_n \\ 0 \\ \vdots \\ 0 \\ \cos\phi_n \end{pmatrix}\begin{pmatrix} -\sin\phi_n \\ 0 \\ \vdots \\ 0 \\ \cos\phi_n \end{pmatrix}^T \frac{1}{\cos\phi_n}\right]$$

(recall that $\phi_n = 0$ so the first summand drops out and by the product rule this

equals to)

$$
\begin{aligned}
&= \frac{-2\sin 2\phi_n}{\rho\cos^2 2\phi_n}\begin{pmatrix}\rho\sin\phi_n\\0\\\vdots\\0\\\cos\phi_n\end{pmatrix}\begin{pmatrix}-\sin\phi_n\\0\\\vdots\\0\\\cos\phi_n\end{pmatrix}^T\\
&+ \frac{1}{\rho\cos 2\phi_n}\left(\begin{pmatrix}\rho\cos\phi_n\\0\\\vdots\\0\\-\sin\phi_n\end{pmatrix}\begin{pmatrix}-\sin\phi_n\\0\\\vdots\\0\\\cos\phi_n\end{pmatrix}^T + \begin{pmatrix}\rho\sin\phi_n\\0\\\vdots\\0\\\cos\phi_n\end{pmatrix}\begin{pmatrix}-\cos\phi_n\\0\\\vdots\\0\\-\sin\phi_n\end{pmatrix}^T\right)\\
&= \frac{1}{\rho}\begin{pmatrix}\rho\\0\\\vdots\\0\\0\end{pmatrix}\begin{pmatrix}0\\0\\\vdots\\0\\1\end{pmatrix}^T + \frac{1}{\rho}\begin{pmatrix}0\\0\\\vdots\\0\\1\end{pmatrix}\begin{pmatrix}-1\\0\\\vdots\\0\\0\end{pmatrix}^T\\
&= \begin{pmatrix}0&0&\dots&0&1\\0&0&\dots&0&0\\&&\vdots&&\\0&0&\dots&0&0\\0&0&\dots&0&0\end{pmatrix} + \begin{pmatrix}0&0&\dots&0&0\\0&0&\dots&0&0\\&&\vdots&&\\0&0&\dots&0&0\\-\frac{1}{\rho}&0&\dots&0&0\end{pmatrix}
\end{aligned}
$$

Multiplied by f, as $\rho \downarrow 0$, to get the last component of the last column of the Jacobian possibly non-zero, this gives us

$$\lim_{\rho\downarrow 0} -\frac{1}{\rho}(f_1) = \lim_{\rho\downarrow 0}\frac{\sum_{i=1}^{n+1}\frac{\partial f_1}{\partial x_i}\frac{\partial x_i}{\partial \rho}}{-1} = \lim_{\rho\downarrow 0}\frac{\frac{\partial f_1}{\partial x_1}\frac{\partial x_1}{\partial \rho}}{-1} = -\frac{\partial f_1}{\partial x_1}$$

Thus, finally, we can write

$$\nabla_{\rho,\phi}\widetilde{f}(0,\mathbf{0}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & & & & \\ & (\frac{\partial f_2}{\partial x_2} - \frac{\partial f_1}{\partial x_1}) & & & \\ & & (\frac{\partial f_3}{\partial x_3} - \frac{\partial f_1}{\partial x_1}) & & \\ & & & \ddots & \\ & & & & (\frac{\partial f_{n+1}}{\partial x_{n+1}} - \frac{\partial f_1}{\partial x_1}) \end{pmatrix}$$

BIBLIOGRAPHY

[1] V. Chvatal, *Linear Programming, W. H. Freeman & Company, 1983*

[2] Ben-Tal, A. and Nemirovski, A., *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications. SIAM-MPS Series in Optimization, 2001*

[3] J. Lasserre, *Global optimization with polynomials and the problem of moments, SIAM Journal of optimization, 11 (2001)*

[4] D. Bertsimas and I. Popescu, *On The Relation Between Option And Stock Prices: An Optimization Approach, Operations Research 50, No. 2 pp. 358-374, March-April 2002*

[5] Jun Sun, Stephen Boyd, Lin Xiao, and Persi Diaconis, *The Fastest Mixing Markov Process on a Graph and a Connection to a Maximum Variance Unfolding Problem, submitted to SIAM Review, problems and techniques section, May 2004*

[6] Yu. Nesterov, Arkadii Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming. SIAM studies in applied mathematics; vol. 13, 1993*

[7] J. Renegar, *A Mathematical View of Interior-Point Methods in Convex Optimization. SIAM-MPS Series in Optimization, 2001*

[8] J. Renegar, *Hyperbolic Programs, and Their Derivative Relaxations, Optimization Online, March 2004*

[9] A.S. Lewis, P.A. Parrilo, M.V. Ramana, *The Lax conjecture is true, Optimization Online, April 2003, submitted to Proceedings of the American Mathematical Society*

[10] H. Bauschke, O. Guler, A.S. Lewis and H.S. Sendov, *Hyperbolic polynomials and convex analysis, Canadian Journal of Mathematics 53 (2001), 470-488*

[11] O. Güler, *Hyperbolic polynomials and interior point methods for convex programming, Mathematics of Operations Research 22 (1997) 350-377*

[12] Lars Gårding, *An inequality for hyperbolic polynomials, J. Math. Mech. 8 (1959), 957-965*

[13] R. Benedetti, *Real algebraic and semi-algebraic sets (Actualite's mathe'matiques), Hermann, 1990*

[14] Chek Beng Chua, *Relating Homogeneous Cones and Positive Definite Cones via T-Algebras, SIAM Journal on Optimization Volume 14, Number 2 pp. 500-506*

[15] J. William Helton, Victor Vinnikov, *Linear Matrix Inequality Representation of Sets, June 2003*

[16] Ben-Tal, A. and Nemirovski, A., *On Polyhedral Approximations of the Second-Order Cone, Mathematics of Operations Research 26 (2001) 193 -205*

[17] M. Chu, Yu. Zinchenko, S.G. Henderson, M.B. Sharpe, *Robust IMRT treatment planning, 2004, working paper*

[18] Morris W. Hirsch, Stephen Smale, *Differential Equations, Dynamical Systems, and Linear Algebra. Academic Press, Inc., 1974*

[19] Lenore Blum, Felipe Cucker, Michael Shub, Steve Smale, *Complexity and Real Computation. Springer-Verlag New York, Inc., 1998*

[20] Vladimir I. Arnold, *Ordinary Differential Equations. The MIT Press, 1978*

[21] Peter Lancaster, *Theory of Matrices. Academic Press, 1969*

[22] Dmitri Alekseevsky, Andreas Kriegl, Mark Losik, Peter W. Michor, *Choosing roots of polynomials smoothly, Israel Journal of Mathematics 105, 1998, 203-233*

[23] P. A. Parrilo, *Semidefinite programming relaxations for semialgebraic problems, Mathematical Programming A, 2002*

[24] N. Z. Shor, *Class of global minimum bounds of polynomial functions, Cybernetics, vol. 23, no. 6, 1987, 731734*

[25] R. E. Curto and L. A. Fialkow, *The Truncated Complex K-moment Problem, Transactions of the American Mathematical Society, Vol. 352, 2000, 28252855*

[26] Y. Nesterov, *Squared functional systems and optimization problems, High Performance Optimization, Kluwer Academic Publishers, 2000, 405440*